Universidad Carlos III de Madrid

Escuela Politécnica Superior.

Grado en Sistemas Audiovisuales

Trabajo Fin de Grado

RECONOCIMIENTO DE EMOCIONES A PARTIR DE IMAGEN Y VOZ

Autor: Diego Sánchez Angón

Tutor: Ascensión Gallardo Antolín

Agradecimientos

A mis padres, por todo el esfuerzo que han hecho para ayudarme a llegar hasta aquí,

A María, por su cariño y su inestimable apoyo durante todo este camino,

A Jorge, porque junto a él he aprendido más que con ningún profesor,

A Ascen, por su esfuerzo y dedicación como tutora de este trabajo.

Gracias de corazón.

Resumen

La interacción entre seres humanos y máquinas ya no es cosa de ciencia ficción, es un hecho cotidiano en la vida de millones de personas en todo el mundo. El objetivo de este trabajo es implementar un sistema de reconocimiento de emociones a partir de imagen y voz, con la motivación de hacer de esta interacción algo más humano y natural.

Para ello, primeramente, se ha estudiado la naturaleza de las emociones, las diferentes teorías existentes para clasificarlas y su influencia en la expresión facial y en la voz de las personas. Más tarde, se han seleccionado diferentes técnicas de Aprendizaje Automático para conseguir que la máquina aprenda por sí misma a reconocer patrones a partir de ejemplos debidamente etiquetados. Además, se identificaron los diferentes tipos de bases de datos existentes relacionadas con las emociones, comparando sus ventajas e inconvenientes.

El sistema resultante posee los bloques básicos de cualquier sistema de clasificación: base de datos, extracción de características y clasificador. La base de datos utilizada consta de 4 locutores, 7 emociones (6 básicas + neutra) y 15 enunciados por emoción (30 en el caso de la neutra). Esto hace un total de 120 enunciados por locutor grabados tanto en vídeo como en audio. Las características en imágenes consisten en la media y la desviación típica de las coordenadas de 60 marcas pintadas en la cara de los locutores durante el período de grabación. Las características de la voz consisten en los coeficientes mel-cepstrales y la frecuencia fundamental. Los clasificadores utilizados han sido el modelo de mezcla de gaussianas (*Gaussian Mixture Model*, GMM) y la máquina de vectores soporte (*Support Vector Machines*, SVM).

En el proceso de implementación, se han generado 6 sistemas de reconocimiento de emociones diferentes, que se clasifican según la naturaleza de los datos empleados (imagen, voz, imagen y voz) y según la dependencia del locutor (dependientes del locutor e independientes del locutor) siendo el sistema definitivo un sistema de reconocimiento de emociones independiente del locutor a partir de imagen y voz.

El resultado de los experimentos llevados a cabo muestra que: los sistemas dependientes del locutor ofrecen mejores prestaciones que los independientes, como era de esperar; los sistemas basados en imagen obtienen mejores resultados que los sistemas basados en voz; los sistemas que fusionan características de imagen y voz mejoran los resultados obtenidos por separado; y el clasificador SVM obtiene una tasa de acierto considerablemente mayor que el GMM.

Abstract

Human-machine interaction is not science fiction anymore; it happens on a daily basis and it concerns millions of people around the world. The aim of this project is developing an emotion recognition system based on image and voice, in order to make this interaction more "human" and natural.

For that purpose, firstly, several factors have been studied: the nature of the emotions, the different theories that classify them and the influence they have in people's facial expression and voice. After that, we have chosen some Machine Learning techniques in order to make the machine learn how to recognize patterns from a collection of examples correctly labelled. Furthermore, we identified different types of databases that contain emotion-related information and we compared their advantages and drawbacks.

The final system contains the basic blocks of any classification system: database, feature extraction and classifier. The database used consists of 4 speakers, 7 emotions (6 basic emotions + neutral) and 15 utterances per emotion (30 for neutral). This makes 120 utterances per speaker recorded in audio and video. Feature extraction from images comprises the mean and standard deviation of the coordinates of 60 markers painted on the speaker's face during the recording. Feature extraction from voice is based on Mel frequency cepstral coefficients and the fundamental frequency. The classifiers that have been employed are the Gaussian Mixture Model (GMM) and the Support Vector Machines (SVM).

In the process of implementation, 6 different emotion recognition systems have been generated. They can be classified according to the nature of the data used (image, voice or image and voice) and according to the speaker dependency (speaker-dependent or speaker-independent) being the ultimate system a speaker-independent emotion recognition system based on image and voice.

The results of the experiments that have been conducted show that: speaker-dependent systems perform better than speaker-independent ones, as was expected; systems based on image get better results than systems based on voice; the fusion of image features with voice features has better results than each one separated; and the SVM classifier achieves a greater recognition rate than the GMM one.

.

Índice General

Capítulo 1

Introducción

En este capítulo se presenta cuál es la motivación del trabajo y se enumeran sus principales objetivos. Además, se hace una breve reflexión sobre el impacto socio-económico del mismo y del marco regulador en el que está inmerso. Por último, se detalla la estructura de esta memoria con la intención de que las personas que se acerquen a ella tengan una mayor facilidad para comprenderla y encontrar en ella lo que buscan.

1.1 Motivación.

Interacción hombre-máquina: de la ficción a la realidad.

En 1968, se entrenaba "*2001: Odisea en el espacio"* de Stanley Kubrick, considerada una de las mejores películas de ciencia ficción de la historia. En esta película, una supercomputadora llamada *HAL 9000* gobierna una nave espacial y se comunica con los humanos mediante el habla gracias a la inteligencia artificial.

Sin embargo, hace cuarenta años, fuera de la gran pantalla, las únicas personas que interactuaban con máquinas eran profesionales del sector de las tecnologías de la información. Sería a finales de los años 70, con la aparición del computador personal (*Personal Computer*, PC) cuando esta interacción empezó a extenderse más allá de los grandes laboratorios informáticos.

Hoy en día, basta con echar un vistazo a nuestro alrededor para darnos cuenta de que no hay una sola persona que no lleve un teléfono móvil inteligente en su bolsillo. Por no hablar de la cantidad de ordenadores, tabletas, consolas de videojuegos, televisiones inteligentes, robots aspiradores y demás dispositivos electrónicos que en los últimos años están invadiendo nuestros hogares. En el año 2017 la interacción entre los

humanos y las máquinas no forma parte de ninguna historia de ciencia ficción, sino de nuestro día a día.

Para ayudar a entender la magnitud de este fenómeno se comentan a continuación algunas cifras. Según un estudio de la consultora y analista Gartner, 2016 cerró con más de 1.500 millones de "smartphones" comercializados [1]. En ese mismo año, según el informe "Mobility Report" de Ericsson, el número de líneas móviles alcanzaba la cifra de habitantes mundiales [2]. Otro informe publicado recientemente por la misma consultora estima que a finales de 2017 el número de dispositivos conectados a internet será de 8.400 millones [3]. Por último, según un informe realizado por "Statista", España está entre los 10 países que más usan el teléfono móvil, con una media de 2 horas y 11 minutos al día [4].

Esta interacción hombre-máquina no siempre se ha producido de la misma manera. Los primeros ordenadores únicamente entendían una serie de comandos que el usuario tenía que introducir mediante el uso de tarjetas perforadas y, más adelante, mediante el teclado. Más tarde, se fue perfeccionando. La llegada de los computadores personales a los hogares de la gente común dejó patente la falta de accesibilidad y usabilidad de los ordenadores. Fue entonces cuando empezaron a aparecer las primeras interfaces gráficas de usuario (*Graphical User Interface*, GUI). Desde entonces, los avances en el campo de la interacción hombre-máquina no han dejado de sucederse y todos parecen encaminados a hacer de esta interacción algo más natural y "humano". Prueba de ello son los avances en tecnologías como el procesado de lenguaje natural (*Natural Language Processing*, NLP) o la síntesis de habla (*Text-To-Speech*, TTS), además de tecnologías como la visión por computador que ha visto multiplicadas sus aplicaciones junto con otras ciencias como la kinesiología o la biometría.

Este trabajo se enmarca dentro de este campo de investigación que busca hacer de la interacción entre personas y máquinas una acción lo más natural posible. Para ello, se pretende investigar el papel que juegan las emociones a la hora de comunicarnos y la posibilidad de dotar a las máquinas de la capacidad de reconocerlas.

Emociones en la comunicación: no es qué se dice, es cómo se dice [5].

Las emociones importan. La comunicación no verbal revela tanta o más información que las propias palabras, y en ella, las emociones juegan un papel fundamental. Pese a los grandes avances llevados a cabo en los últimos tiempos, las máquinas siguen obviando esta faceta de la comunicación humana. Tras años de estudio y grandes esfuerzos, hemos conseguido que las máquinas entiendan, no sin dificultad, qué decimos. Sin embargo, parece que no se han destinado los mismos esfuerzos a que comprendan cómo lo decimos.

En la década de las redes sociales y los chats de mensajería instantánea, no hace falta explicar las limitaciones que tiene la comunicación cuando carece de su carácter no verbal. Estas aplicaciones están llenas de emoticonos (que no es más que un neologismo

que proviene de emoción e icono) para intentar paliar esta ausencia de lenguaje no verbal en el texto escrito. Como el lector seguramente habrá comprobado, los malentendidos y las frases sacadas de contexto son constantes.

De hecho, según el psicólogo alemán Albert Mehrabian [6], sólo el 7 por ciento de la información se atribuye a las palabras, mientras que el 38 por ciento se atribuye a la voz (entonación, intensidad…) y el 55 por ciento a la expresión corporal (gestos, expresiones faciales, mirada…).

La principal motivación de este trabajo es seguir trabajando en que la interacción entre los humanos y las máquinas sea lo más natural posible. Sería deseable que estas máquinas, con las que cada vez interactuamos más, fuesen capaces de tener algún tipo de empatía hacia las personas o de persuadirnos y motivarnos a través de la comprensión de nuestras emociones. Con esta motivación, se abordará el estudio de un sistema de reconocimiento automático de emociones para dotar así a las máquinas de una herramienta fundamental para interaccionar más y mejor con los seres humanos.

1.2 Objetivos.

El objetivo fundamental del trabajo es la realización de un sistema de reconocimiento automático de emociones a partir de imágenes y voz, o lo que es lo mismo, vídeos. Este sistema debe ser capaz de etiquetar con la emoción correspondiente un conjunto de vídeos procedentes de una base de datos especialmente creada para el reconocimiento de emociones.

Además de este objetivo principal, otros objetivos secundarios son:

- Estudiar la naturaleza de las emociones y el origen de las mismas, además de investigar las diferentes teorías sobre cómo se clasifican.
- Estudiar las bases del aprendizaje automático, identificar sus principales tareas y, en concreto, entender el problema de la clasificación.
- Investigar la expresión vocal de las emociones, identificando cuáles son las principales características de la voz que nos permiten reconocer las emociones.
- Investigar la expresión facial de las emociones, averiguar qué son las unidades de acción y estudiar cómo se pueden codificar las emociones a partir de ellas.
- Identificar los diferentes tipos de bases de datos existentes para el reconocimiento de emociones y entender sus principales ventajas e inconvenientes.
- Analizar los resultados obtenidos y evaluar si la fusión de la parte auditiva con la visual mejora las prestaciones del sistema respecto a la utilización de cada una por separado.
- Generar conclusiones al respecto e identificar cuáles serían las posibles líneas futuras para seguir investigando en este campo.

1.3 Impacto socio-económico.

Las emociones no mienten y, además, son muy difíciles de ocultar. Esto, añadido a todo lo comentado en el apartado 1.1, hace que un sistema de reconocimiento de emociones sea una herramienta realmente valiosa en diferentes ámbitos y sectores, y que sus aplicaciones sean tan numerosas como variadas. A continuación, se describe el impacto socio-económico de algunos ejemplos de dichas aplicaciones, según el ámbito o sector en el que se enmarcan [7]:

Medicina y Psicología.

- Evaluar el dolor que sufren los niños con ciertos tratamientos.
- Ayudar a los niños con autismo en el reconocimiento de emociones a partir de las expresiones faciales.
- Evaluar el progreso en procesos de rehabilitación en niños o personas mayores.
- Ayudar a los psicólogos a detectar el estado emocional de los pacientes

Educación.

- Obtener "feedback" en tiempo real de los alumnos ante explicaciones o ejercicios. Esto permite detectar qué seguimiento está teniendo la clase por parte de cada alumno y tomar medidas al respecto.

Marketing, publicidad y retail.

- Analizar el recibimiento por parte de los consumidores de ciertas campañas de marketing.
- Personalizar la oferta de productos en tiempo real tras analizar las emociones de los clientes, por ejemplo, en sitios web.
- Analizar la reacción de los clientes al observar productos en el escaparate de una tienda.

Política y demoscopia.

- Analizar la reacción de la gente ante campañas políticas, ya sea en mítines, en los vídeos lanzados durante la campaña o en los grandes debates.
- Analizar las emociones en los discursos de políticos para saber si realmente sienten lo que dicen.
- Analizar las emociones en respuestas a encuestas para evitar respuestas falsas.

Entretenimiento.

- Para distribuidores de contenido multimedia, permitiría monitorizar las emociones de los espectadores del contenido para después analizar qué parte les ha gustado más, o si se han conseguido transmitir las emociones deseadas en cada momento.

- Personalizar la experiencia en videojuegos a través del análisis en tiempo real de las emociones del jugador.

1.4 Marco regulador.

Uno de los grandes problemas de todas las aplicaciones mencionadas en el apartado anterior es el problema de la privacidad y de la protección de los datos. Es necesaria una legislación para abordar los problemas con la privacidad de las personas que puedan derivarse de este tipo de aplicaciones.

En España existe una ley de protección de datos: la Ley Orgánica 15/1999 de 13 de diciembre de Protección de Datos de Carácter Personal (LOPD) [8].

El 25 de Mayo de 2016 la Unión Europea aprobó el Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos [9]. Se prevé que para el año 2018 se apruebe en España una modificación de la LOPD para cumplir con este reglamento europeo.

Para la creación de bases de datos audiovisuales, como la usada en este proyecto, las personas que aparecen en ellas deben firmar un documento para dar permiso para la grabación y distribución de sus datos biométricos.

En lo que respecta a la realización de nuestro proyecto, la base de datos utilizada cuenta con el consentimiento de las personas que en ella aparecen, para ser utilizada de manera libre para fines académicos.

1.5 Estructura de la memoria.

Con la finalidad de facilitar al lector el acercamiento y compresión del presente trabajo, se detalla a continuación un breve resumen de los capítulos que lo componen:

En el capítulo 1 se expone la motivación y los objetivos del trabajo, además de su impacto socio-económico y marco regulador. Finalmente, se hace un breve resumen de la estructura del mismo y se repasa el trabajo previo en el que se basa.

En el capítulo 2 se hace un estudio detallado del estado del arte en el que se enmarca el proyecto. Primeramente, se estudia el concepto de emoción y enumeran las diferentes teorías para clasificarlas, desde el punto de vista de la Psicología. Seguidamente, se exponen las principales características de la expresión, tanto vocal como facial, de las emociones. Tras ello, se hace una introducción al aprendizaje automático. Finalmente, se explican los diferentes tipos de bases de datos que existen para el reconocimiento de emociones, identificando sus principales virtudes y defectos.

En el capítulo 3 se explica cómo se ha llevado a cabo la implementación del sistema de reconocimiento de emociones. En concreto, se describen la base de datos, las características, tanto vocales como faciales, y los clasificadores empleados.

En el capítulo 4 se analizan los experimentos llevados a cabo para evaluar las prestaciones de los diferentes sistemas y se extraen conclusiones a partir de sus resultados.

El capítulo 5 aborda la parte de gestión del proyecto, explicando la planificación del mismo y detallando el presupuesto.

En el capítulo 6 se explican las conclusiones extraídas del trabajo realizado y se presentan posibles líneas futuras de investigación.

1.6 Trabajo previo.

Este trabajo se basa en el trabajo previo realizado por el *Centre of Vision, Speech and Signal Processing* (CVSSP) de la Universidad de Surrey en Reino Unido. Concretamente, este trabajo se basa en el documento publicado por Sanaul Haq y Philip J.B. Jackson de título "*Speaker-Dependent Audio-Visual Emotion Recognition*" [10]. Además, se ha utilizado para la realización de este trabajo la base de datos creada por dicha universidad, *Surrey Audio-Visual Expressed Emotion (SAVEE) Database* [11].

Capítulo 2

Estado del arte

Este capítulo trata de hacer una radiografía del contexto tecnológico en el que se produce este proyecto y de dar fundamento teórico a la implementación del mismo. Primeramente, se presenta el concepto de emoción y se comentan las diferentes teorías sobre la clasificación de emociones. Después, se hace una introducción al aprendizaje automático. Seguidamente, se explican las diferentes características que tienen la expresión vocal y la expresión facial de las emociones. Finalmente, se identifican los diferentes tipos de bases de datos utilizadas para el reconocimiento de emociones, con sus ventajas e inconvenientes.

2.1 Las emociones.

Una vez han quedado claros la motivación y el objetivo de este trabajo (la realización de un sistema de reconocimiento de emociones para hacer de la interacción humano-máquina algo más natural), parece sensato comenzar este trabajo por investigar qué es aquello que queremos reconocer: las emociones.

En el apartado 1.1 ya se ha comentado que las emociones juegan un papel fundamental en el proceso de comunicación de los seres humanos. Pero, más allá de la comunicación, las emociones influyen constantemente en nuestra forma de pensar y de actuar. En nuestro día a día tomamos decisiones que están claramente afectadas por nuestras emociones: si estamos contentos o tristes, si tenemos miedo o si estamos furiosos. De esta manera, entender la naturaleza de las emociones, qué tipos de emociones existen o cuál es el origen de las mismas, parecen, a todas luces, cuestiones primordiales para entender nada más y nada menos que la conducta del ser humano.

Etimológicamente, la palabra emoción viene del latín *emotio*, que significa "movimiento o impulso", "aquello que te mueve hacia". En el campo de la psicología, numerosos investigadores, a lo largo de las últimas décadas (incluso siglos), se han

enfrentado a la difícil tarea de definir qué son las emociones y cuál es el origen de las mismas. Fruto de sus estudios, han surgido diferentes teorías al respecto [12].

Las teorías evolutivas, como la que Darwin expuso en su libro *The Expression of the Emotions in Man and Animals* (1872) [13], argumentan que las emociones son el resultado del propio proceso evolutivo y que han permitido, tanto a humanos, como a animales, adaptarse al medio para sobrevivir y reproducirse. Según esta teoría, que un animal sienta miedo ante situaciones de amenaza sería una ventaja evolutiva que les ha permitido sobrevivir a este tipo de situaciones. Además, como consecuencia de lo anterior, los defensores de estas teorías están de acuerdo en la existencia de ciertas emociones universales, independientes del contexto socio-cultural.

Las teorías fisiológicas como la de James-Lange entienden que el origen de cualquier emoción es una excitación fisiológica y que sólo después de esta excitación experimentamos la emoción. De esta manera, si no sintiésemos la aceleración de nuestro pulso o la segregación de sudor en ciertas partes de nuestro cuerpo, no experimentaríamos la emoción del miedo.

Las teorías neurológicas como la de Cannon-Bard interpretan que experimentamos la excitación fisiológica y la emocional al mismo tiempo.

Las teorías cognitivas como la de Schachter-Singer, creen también, como las fisiológicas, que lo primero que experimentamos es una excitación fisiológica, pero además, creen que la experimentación de la emoción es fruto de la interpretación que da el cerebro a esta excitación.

Independientemente de su origen, en lo que sí parecen coincidir la mayoría de los psicólogos es en que, como escribió el matrimonio Hockenbury en su libro *Descubriendo la Psicología* [14], "una emoción es un estado psicológico complejo que implica tres componentes distintos: una experiencia subjetiva, una respuesta fisiológica, y una respuesta conductual o expresiva".

Basándonos en esta definición podemos distinguir tres componentes principales que conforman una emoción [15]:

- La experiencia subjetiva, que es lo que a veces denominamos sentimiento. Sentimos miedo, ira, asco…
- La respuesta fisiológica, que tiene que ver con la reacción del cuerpo que se manifiesta en respuestas como taquicardia, sudoración, sequedad en la boca…
- La respuesta conductual o expresiva, que tiene que ver con el comportamiento del individuo que se manifiesta en las expresiones faciales, el tono de voz, el volumen, los gestos…

Para la realización de este trabajo, la componente de las emociones que más nos interesa estudiar es la respuesta conductual o expresiva, ya que es la que en última instancia permite inferir el estado emocional de una persona a partir de la observación de su expresión facial y de las características de su voz.

2.1.1 Clasificación de emociones [16]

Si el problema de definir qué son las emociones y de identificar su origen es realmente complicado, la tarea de clasificarlas no lo es menos. El componente subjetivo de las emociones las hace difíciles de clasificar. ¿Todas las personas se enfadan de la misma manera? ¿Todas las personas reaccionan emocionalmente igual ante un mismo hecho? La respuesta es que no, y es exactamente ahí donde radica la dificultad de encontrar una clasificación de emociones universal, que valga para cualquier persona.

Este problema no es nuevo. Ya en el siglo IV a.c. el famoso filósofo griego Aristóteles intentó identificar el número exacto de emociones fundamentales que existen. Llegó a la conclusión de que existen 14 emociones irreductibles, entre las que se encontraban el miedo, el enfado, la vergüenza o la pena, entre otras.

Más recientemente, varios psicólogos han tratado de dar respuesta a este problema, surgiendo también diferentes teorías respecto a la clasificación de las emociones. A continuación se explican algunas de ellas.

Emociones básicas.

Uno de los psicólogos con mayor reputación en el campo de las emociones es Paul Ekman. En 1972, tras su viaje a Papúa Nueva Guinea, llegó a la conclusión de que existían 6 emociones básicas y universales [17]. En la Tabla 1 se muestran estas 6 emociones:

Emociones básicas
Alegría
Tristeza
Sorpresa
Asco
Miedo
Ira

Tabla 1. Emociones básicas de Paul Ekman.

Emociones positivas y negativas.

Otro tipo de clasificación muy común es la que diferencia entre emociones positivas, negativas y neutras (ver Tabla 2).

Positivas	Negativas	Neutras
Alegría Amor Felicidad Confianza Diversión Satisfacción	Ira Miedo Asco Arrepentimiento Tristeza Rabia	Sorpresa Esperanza Compasión

Tabla 2. Ejemplos de emociones positivas, negativas y neutras.

Emociones primarias y secundarias.

Por otro lado, algunos autores han hablado de emociones primarias y emociones secundarias, de manera que las segundas se desprenden de las primeras. Así, la alegría, por ejemplo, sería una emoción primaria, mientras que la diversión, la euforia o el entusiasmo serían emociones secundarias que se desprenden de la alegría. La Tabla 3 muestra un ejemplo de este tipo de clasificación.

Primarias	Secundarias
Ira Alegría Miedo Tristeza Amor Sorpresa Vergüenza Asco	Furia Odio Diversión Satisfacción Terror Simpatía Desprecio Humillación Etc…

Tabla 3. Ejemplos de emociones primarias y secundarias

Otro ejemplo de este tipo de clasificaciones es la rueda de las emociones de Robert Plutchik (ver Figura 1). Este tipo de clasificación consiste en dividir las emociones en 8 categorías básicas (primarias) de las cuales se derivan otras emociones (secundarias) según el grado de intensidad.

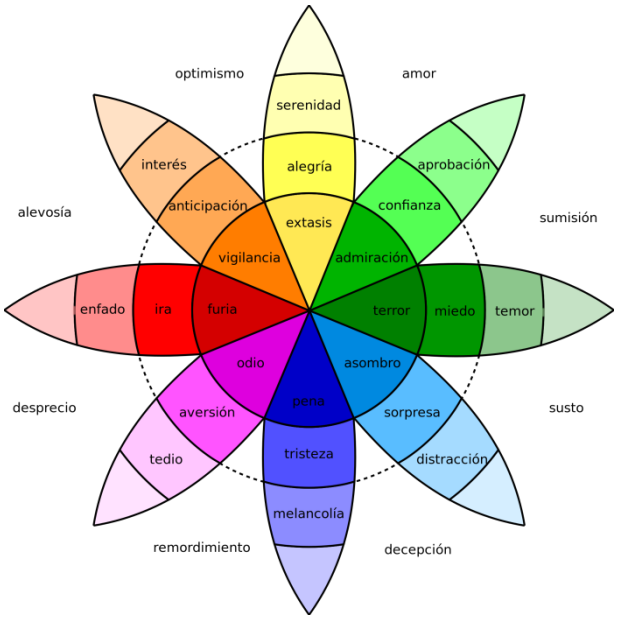


Figura 1. Rueda de las emociones de Robert Plutchik [18].

Emociones en el espacio tridimensional.

Existen otro tipo de teorías que consideran que las emociones pueden representarse en un espacio tridimensional. La principal diferencia entre ellas es la elección de estas dimensiones. La Figura 2 y la Figura 3 son dos ejemplos de espacios tridimensionales.

El modelo de estado emocional PAD (**P**leasure, **A**rousal, **D**ominance), desarrollado por Albert Mehrabian y James A. Russell [19], sugiere que las tres dimensiones que definen este espacio son:

- **Placer** (Pleasure)
- **Activación** (Arousal)
- **Dominio** (Dominance)

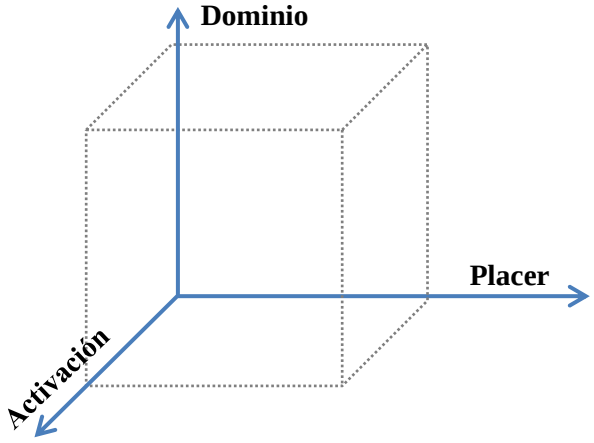


Figura 2. Modelo de estado de emocional (PAD) de A.Mehrabian y J.A Rusell

El sistema propuesto por Joel Davitz y Klaus Scherer [20] utiliza las tres dimensiones de un campo semántico, es decir, están relacionadas con diferentes características de la voz:

- **Potencia** o fuerza: corresponde con el ritmo, las inflexión o la entonación
- **Valencia** o **agrado**: evalúa lo placentero o desagradable de la emoción.
- **Actividad**: se refiere a la velocidad, timbre y tono de la voz.

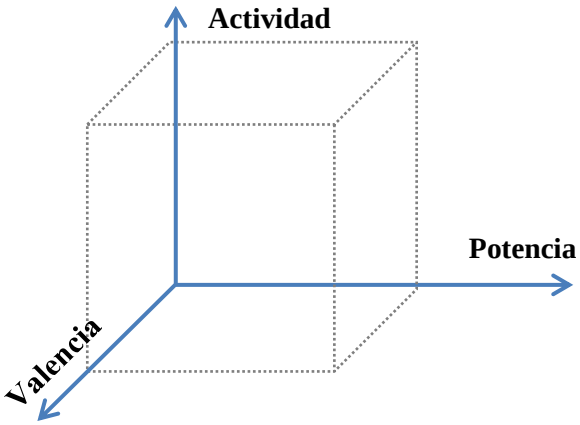


Figura 3. Modelo basado en las tres dimensiones de un campo semántica de J.Davitz y K.Scherer

2.2 Expresión vocal de las emociones.

La voz humana es y ha sido a lo largo de la historia el principal medio que han tenido los humanos para comunicarse. Además de la información implícita en el propio mensaje, la voz contiene gran cantidad de información sobre la persona que lo está emitiendo, como su edad, su género, su origen, su cultura o incluso su manera de ser. Además, como no podía ser de otra forma, la voz es una herramienta fundamental para el reconocimiento de emociones.

La voz es un sonido producido por el aparato fonador humano y, como tal, posee las siguientes características fundamentales:

- **Intensidad**. Esta cualidad está relacionada con la amplitud de la onda sonora y la energía que ésta transporta. Nos permite clasificar los sonidos en sonidos fuertes o débiles.

- **Tono**. Esta cualidad del sonido está relacionada con su frecuencia. Nos permite clasificar los sonidos en sonidos agudos (alta frecuencia) o graves (baja frecuencia).

- **Duración**. Esta cualidad tiene que ver con el tiempo que se prolonga el sonido, o lo que es lo mismo el tiempo que perdura la fuente que lo provoca. Nos permite clasificar los sonidos en sonidos cortos o largos.

- **Timbre**. Esta cualidad está relacionada con la forma de la onda y permite distinguir dos sonidos de la misma intensidad y mismo tono.

Sin embargo, estas tres características no son de gran utilidad para la implementación de nuestro sistema de clasificación, ya que son características subjetivas y con un enfoque cualitativo, es decir, se basan en cómo es percibido el sonido por el oyente y no se pueden cuantificar o medir. Para llevar a cabo este sistema es necesario extraer características más objetivas y con un enfoque cuantitativo, esto es, características que no dependan tanto del oyente y que puedan ser medidas.

Existen numerosas características de la voz que cumplen con estos requisitos. Entre ellas podemos distinguir varios tipos:

- **Características prosódicas**. Estas características, también conocidas como suprasegmentales, son aquellas que tienen que ver con elementos que están un nivel por encima de los segmentos fonéticos (fonemas), como las sílabas, las palabras o secuencias de palabras. Entre estas características se encuentran:
  - La frecuencia fundamental.
  - El ritmo.
  - La intensidad.

- **Características espectrales**. Estás características tratan de representar la distribución de la energía entre las diferentes frecuencias de la voz. Entre ellas se encuentran:
  - Densidad espectral de energía.
  - Coeficientes mel-cepstrales (*Mel-Frequency Cepstral Coefficients*, MFCC)
  - Transformada ondicular (*wavelet*).

- **Calidad de la voz.** Este tipo de características representan la influencia que tiene la forma en que está configurado el aparato vocal del locutor en el sonido de su voz. La mayoría de las veces estas características están relacionadas con elementos que se encuentran en el interior de la laringe, como las cuerdas vocales o la glotis. Entre ellas se encuentran:
  - *Jitter*.
  - *Shimmer*.
  - El cociente entre energía de alta y baja frecuencia.

2.3 Expresión facial de las emociones.

Se suele decir que los ojos son el espejo del alma, y en cierto modo, es cierto. La expresión facial de las personas es un fiel reflejo de su estado emocional.

Desde Darwin hasta la actualidad, la expresión facial de las emociones ha supuesto un campo de estudio fascinante. Autores como Tomkins, Izar y Ekman han asumido el carácter innato y universal de las acciones faciales que son reflejo de las emociones.

Según estos autores, la expresión facial de las emociones es universal y no depende de ningún contexto social o cultural, contradiciendo así la teoría tradicional. En uno de sus viajes Ekman demostró que incluso en tribus alejadas de todo contacto con la civilización, las personas eran capaces de reconocer correctamente qué emociones representaban ciertas expresiones faciales que se les mostraban en fotografías [17].

Uno de los retos de estos autores ha sido la codificación de las emociones en unidades de acción faciales, de manera que se pueda determinar la emoción que está expresando una cara, únicamente conociendo cuáles de estas unidades están activas o no. En 1978, Paul Ekman y Wallace V. Friesen publicaron un sistema de codificación facial basado en un sistema desarrollado por un anatomista sueco llamado Carl-Herman. Más tarde, en 2002, Ekman, Friesen, and Joseph C. Hager publicaron una importante actualización.

Sistema de codificación facial (FACS)

El Sistema de Codificación Facial (*Facial Action Coding System*, FACS) [21] es un sistema para medir todos los movimientos faciales distinguibles. El FACS describe todos estos movimientos basándose en 44 unidades de acción (*Action Unit*, AU) únicas, además de en otras categorías que representan el movimiento y la posición de la cabeza y los ojos. Estas unidades de acción representan la contracción de distintos músculos de la cara. De esta manera, con este sistema, la descripción de una expresión facial, conocida como evento, consiste en una lista de las unidades de acción activas en dicha expresión. En la Figura 4 podemos ver listadas las 27 primeras unidades de acción, junto con una explicación y su expresión facial correspondiente.

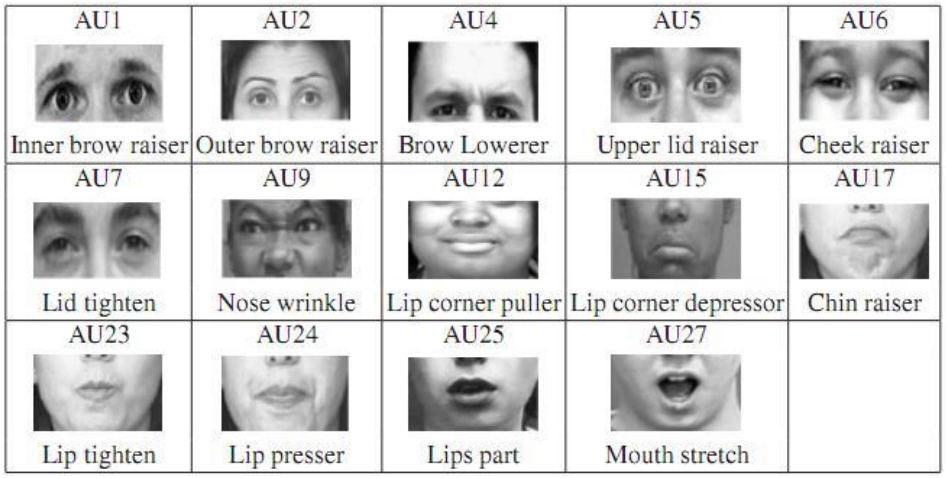


Figura 4. Tabla con las primeras 27 *Action Units* con su explicación y expresión facial correspondientes [22].

2.4 Aprendizaje automático [23].

Nadie nace sabiendo. Los seres humanos venimos a este mundo sin ningún conocimiento previo. Todos los conocimientos que adquirimos a lo largo de nuestra vida se los debemos al aprendizaje. Desde que nacemos, el aprendizaje actúa como fuente continua de mejora.

Desde el campo de la psicología, el concepto de aprendizaje encierra numerosas incógnitas. En realidad, se puede afirmar que no se conocen de manera fehaciente los mecanismos y engranajes que manejan el proceso de aprendizaje en los seres humanos. Sin embargo, sí que existen definiciones ampliamente reconocidas por los investigadores del campo de la inteligencia artificial del concepto de aprendizaje, basándose en su función más que en su naturaleza. Algunas de estas definiciones son las siguientes:

- *Aprender es construir o modificar representaciones de aquello con lo que se está experimentando.* [McCarthy, 1968]
- *El aprendizaje denota cambios en el sistema que permiten que se realice la misma tarea más eficiente y eficazmente la próxima vez.* [Simon, 1983]
- *Aprender es hacer cambios útiles en nuestra mente.* [Minsky, 1985]

El aprendizaje automático (*machine learning* en inglés) es una rama de la inteligencia artificial que nace de la constatación de que es imposible construir sistemas que aparenten ser inteligentes, si estos sistemas no poseen la capacidad de adquirir conocimientos o de mejorar su comportamiento a partir de la experiencia, esto es, en definitiva, de aprender. De manera que el principal objetivo del aprendizaje automático es dotar a las máquinas de la capacidad de aprender. La pregunta es, ¿es posible hacer que las máquinas aprendan? A día de hoy, podemos responder a esta pregunta con una rotunda afirmación. El aprendizaje automático ya es una realidad en un montón de

aplicaciones como traductores automáticos, coches autónomos, biomedicina, concesión de créditos, control del tráfico, sistemas de recomendación…etc.

Sin embargo, la forma de aprender que tienen las máquinas no es igual que la que tienen los humanos, aunque sí guarden similitudes. En el proceso de aprendizaje de las máquinas se distinguen claramente dos fases, las cuales es muy importante diferenciar: la fase del aprendizaje en sí, también conocida como fase de entrenamiento, y la fase de utilización de ese conocimiento aprendido, también conocida como fase de prueba.

- **Fase de entrenamiento**, a partir de ejemplos u observaciones previas se genera un modelo o función de discriminación.
- **Fase de prueba**, se comprueba las prestaciones del modelo con nuevos ejemplos u observaciones no utilizados en la fase de entrenamiento.

Además, el proceso de aprendizaje de las máquinas puede tomar diferentes formas, según la naturaleza de la tarea que queramos aprender. Así pues, se puede distinguir entre tres tipos de aprendizaje: supervisado, no supervisado y mixto. En este trabajo, únicamente comentaremos brevemente el aprendizaje supervisado, que es el que utilizamos en nuestro sistema de reconocimiento.

El aprendizaje supervisado es aquel en el que los ejemplos utilizados en la fase de entrenamiento están debidamente etiquetados, estos es, clasificados como positivos o negativos en relación a la tarea que se quiere aprender.

Dentro de los problemas de aprendizaje supervisado, una de las tareas tipo más representativas es la tarea de clasificación. Esta tarea se caracteriza porque las clases, esto es, los valores de las cosas que queremos aprender, se conocen con anterioridad. Además, se caracteriza porque el resultado de esta tarea es una función discriminatoria entre dichas clases. Un tipo especial de clasificación es la conocido como decisión, que no es más que una clasificación binaria entre ejemplos positivos y negativos. Nuestro sistema de reconocimiento de emociones es un claro ejemplo de la tarea de clasificación en el cual las observaciones son vídeos de personas expresando diferentes emociones, las clases son diferentes tipos de emociones,, y el resultado debe de ser una función que discrimine estos vídeos según pertenezcan a una u otra emoción.

2.4.1 Descripción del proceso de aprendizaje.

El proceso llevado a cabo por una máquina en un problema de aprendizaje supervisado, es el siguiente:

- **Fase de entrenamiento.**

  Adquisición de instancias u observaciones de entrenamiento.

Es el paso previo indispensable. Esta adquisición de datos se llevará a cabo por la propia máquina mediante sensores como cámaras, micrófonos, acelerómetros, etc. o

serán proporcionados por los humanos mediante bases de datos existentes. Estos datos estarán en crudo, esto es, sin ningún tipo de tratamiento. Según el tipo de sensor estos datos serán de diferente índole. Si, por ejemplo, el sensor es una cámara fotográfica, los datos serán el valor de cada uno de los píxeles de las fotografías tomadas por dicha cámara.

Etiquetado de las instancias de entrenamiento.

Como hemos visto, para realizar tareas de aprendizaje supervisado es necesario tener los ejemplos de entrenamiento correctamente etiquetados. El resultado de este etiquetado será un vector $\mathbf{m_e \times 1}$, donde $\mathbf{m_e}$ es el número de instancias que estamos utilizando para entrenar la máquina.

Extracción de características.

El siguiente paso es extraer características de dichas observaciones, de manera que estas características sean útiles a la hora de llevar a cabo la clasificación. El resultado de este proceso es un vector de características por cada una de las observaciones formando una matriz de tamaño $\mathbf{m_e \times n}$, donde **n** es el número de características extraídas. Este proceso permite representar cada instancia como un punto dentro de un espacio n-dimensional.

Generación del modelo o función discriminatoria.

Una vez hemos conseguido mapear cada instancia de nuestra base de datos a un espacio n-dimensional, la generación del modelo o función discriminatoria no es más que un problema de optimización para encontrar la función que discrimine de la mejor manera posible las instancias según su correspondiente etiqueta. En conjuntos de instancias linealmente separables y n-dimensionales, este proceso se puede ver geométricamente como la búsqueda del hiperplano o conjunto de hiperplanos que dividen el espacio de manera que las instancias con una etiqueta concreta queden en el mismo recinto del espacio. Existen diferentes métodos para llevar a cabo este cometido, los cuales se explicarán en el apartado 2.4.2 dedicado a clasificadores.

- **Fase de prueba.**

1. Adquisición de instancias u observaciones de prueba.

Este paso es prácticamente idéntico al primer paso de la fase de entrenamiento. Sin embargo, es importante tener en cuenta que las instancias utilizadas en la fase de prueba deben ser diferentes de las instancias utilizadas previamente en la fase de entrenamiento. Esto debe ser así para asegurar una buena generalización del modelo. Si no procedemos de esta manera podemos tener lo que se conoce como un problema de sobreajuste, esto es, que nuestro modelo discrimine perfectamente las instancias de entrenamiento, pero que ante muestras nuevas que no haya procesado previamente las prestaciones se vean mermadas.

2. Etiquetado de las instancias de prueba

Al igual que con las instancias de entrenamiento, también es necesario etiquetar correctamente las instancias de prueba. El resultado será un vector $\mathbf{m_p \times 1}$, donde $\mathbf{m_p}$ es el número de instancias de prueba. Esto nos permitirá evaluar las prestaciones del sistema.

Extracción de características.

El siguiente paso será extraer las características de las muestras adquiridas. Al igual que en la fase de entrenamiento obtendremos un vector de características por cada una de las muestras, formando una matriz de tamaño $\mathbf{m_p \times n}$.

Clasificación.

Una vez mapeadas las instancias de prueba al espacio n-dimensional, podemos evaluar la función discriminatoria para cada una de las instancias. Según el tipo de modelo o función discriminatoria el resultado de esta evaluación puede tomar diferentes formas. Por ejemplo, podemos obtener la probabilidad de cada instancia de pertenecer a cada una de las clases. Con esta información, ya se pueden clasificar las instancias en sus correspondientes clases (por ejemplo, eligiendo la clase con la probabilidad más alta de pertenencia para cada muestra). Así, obtendremos un vector de resultados de longitud $\mathbf{m_p x 1}$, que contiene la categoría asignada por el clasificador a cada una de las instancias de prueba.

Evaluación de resultados.

Comparando el vector de etiquetas de las instancias de prueba con el vector de resultados podemos evaluar las prestaciones de nuestro sistema de clasificación. Una manera sencilla de obtener una medida de las prestaciones es dividiendo el número de aciertos entre el número total de instancias, lo que nos da el porcentaje de acierto de nuestro sistema.

2.4.2 Clasificadores

Como hemos visto en el apartado anterior, los problemas de aprendizaje supervisado se basan en la creación de un modelo o función discriminatoria a partir de una serie de ejemplos debidamente etiquetados. Para la creación de estos modelos existen una gran variedad de algoritmos, también conocidos como clasificadores. Para la creación de sistemas de reconocimiento de emociones u otros patrones los más utilizados son los siguientes:

- Gaussian Mixture Model, GMM [24].
- Support Vector Machine, SVM [25].
- Hidden Markov Model, HMM [24], [26].
- Artificial Neural Networks, ANN [27].

2.4.3 Sistema de reconocimiento multimodal.

Uno de los objetivos de este trabajo es evaluar cómo se comporta la fusión de la información proveniente de las imágenes y de las grabaciones de voz a la hora de implementar un sistema de reconocimiento de emociones. Este tipo de sistemas se conocen como sistemas multimodales, ya que utilizan datos de diferente naturaleza o procedencia, esto es, diferentes modalidades. Esta fusión se puede realizar a diferentes niveles:

En la fusión a nivel de características, los vectores de características de cada una de las modalidades son agrupados justo después de ser extraídos.

En la fusión a nivel de decisión, los sistemas de las diferentes modalidades se ejecutan en paralelo hasta el momento de decidir con qué clase o categoría etiquetar una observación. En ese momento se evalúan las salidas de cada uno de los clasificadores y con dicha información se etiqueta la muestra.

2.5 Bases de datos [28].

Como acabamos de explicar, cualquier sistema de aprendizaje automático supervisado necesita de una extensa base de datos debidamente etiquetada para entrenar y probar dicho sistema. En concreto, para realizar un sistema de reconocimiento de emociones basado en imágenes y voz, la base de datos necesaria tendrá que ser de naturaleza audiovisual y deberá representar de manera equilibrada las diferentes emociones que queremos detectar. La creación de bases de datos que representen emociones espontáneas puede presentar ciertos problemas:

- Como comentamos en el apartado 1.4, el marco regulador puede dificultar la obtención de este tipo de bases de datos debido a las leyes de protección de datos.
- La expresión de las emociones tiene un tiempo de vida muy corto y es difícil acotarlo.
- El etiquetado de este tipo de bases de datos puede ser verdaderamente costoso, en términos de tiempo y recursos. Además, es fácil que se cometan errores al llevarlo a cabo.

Por estos motivos la mayoría de bases de datos utilizadas para la creación de sistemas de reconocimiento de emociones son actuadas, no espontáneas. Este tipo de bases de datos se llevan a cabo en laboratorios o entornos totalmente controlados. Allí, con cámaras o micrófonos se graba a un grupo de personas, actores o no, fingiendo las diferentes emociones que queremos detectar. La principal ventaja de este método es que permite un mayor control sobre el diseño de la base de datos.

Sin embargo, las bases de datos actuadas también tienen grandes inconvenientes:

- En general, las emociones en este tipo de bases de datos son más exageradas que en las naturales.
- Los sistemas creados con estas bases de datos son evaluados bajo condiciones óptimas (audio libre de ruido, vídeo de alta calidad, buena iluminación, enfoque...). Sin embargo, en la realidad estas condiciones no se suelen cumplir. Por ello estos sistemas pueden obtener grandes prestaciones con dicha base de datos, y por el contrario, obtener unas prestaciones muy pobres con casos más reales.
- Otro problema tiene que ver con las categorías elegidas para representar las emociones. Normalmente con este tipo de bases de datos se utilizan como categorías emociones básicas. Sin embargo, como ya hemos comentado en el apartado las emociones pueden llegar a tener un grado de complejidad mayor.

Pese a estos problemas, las bases de datos actuadas siguen siendo una buena opción para la creación de un sistema de reconocimiento de emociones, debido al gran control que se tiene sobre el diseño de la misma y a las facilidades que conlleva su uso.

Entre los ejemplos de bases de datos utilizadas para el reconocimiento de emociones encontramos diferentes tipos: algunas únicamente incluyen grabaciones de audio, otras sólo imágenes, mientras que otras son audiovisuales; algunas son naturales, otras actuadas y otras provocadas o suscitadas. En la Tabla 4 se pueden ver algunos ejemplos de ellas.

Nombre	A/V	Método	Tamaño	Categorías
AIBO database (Batliner et al., 2004)	A	Natural: interacción de niños con robots.	110 diálogos	enfadado, aburrido, empático, servicial, irónico, alegre, reprimido, relajado, sorprendido, quisquilloso
Berlin Database (Burkhardt et al., 2005)	A	Actuado	493 oraciones	enfado, aburrimiento, asco, miedo, felicidad, tristeza, neutral
ISL meeting corpus (Burger et al., 2002)	A	Natural	18 reuniones	negativa, positiva, neutral
BU-3DFE database (Yin et al., 2006)	V	Actuado	100 locutores	6 emociones básicas con 4 niveles de intensidad
Cohn-Kanade database (Kanade et al., 2000)	V	Actuado	210 locutores; 480 vídeos	6 emociones básicas
Adult Attachment Interview database (Roisman, 2004)	AV	Natural: entrevistas	60 locutores, entrevistas: 30-60 min.	6 emociones básicas
Busso-Narayanan database (Busso et al., 2007)	AV	Actuado	612 frases; una locutora	enfado, alegría, tristeza, neutral

Tabla 4. Ejemplos de bases de datos utilizadas para el reconocimiento de emociones. Información sacada de [28].

Capítulo 3

Sistema de reconocimiento

En este capítulo se detallan las decisiones de diseño llevadas a cabo para la implementación del sistema de reconocimiento de emociones. Incluye un esquema general del sistema. Además, se indica qué base de datos se ha utilizado, qué métodos han sido empleados para la extracción de características y qué algoritmos de clasificación se han elegido para llevar a cabo dicho sistema.

Como explica el apartado 1.2, el objetivo principal de este trabajo es la implementación de un sistema de reconocimiento de emociones a partir de imagen y voz. Aunque este es el objetivo final del trabajo, en el transcurso de la implementación del sistema final, hemos ido creando diferentes sistemas parciales que nos han servido para evaluar diferentes decisiones de diseño, como las características a extraer o qué tipo de clasificadores utilizar. Aunque estos sistemas parciales no sean el objetivo final del proyecto, sí que se comentará su implementación en este capítulo, además de comentar los experimentos llevados a cabo, sus resultados, y las conclusiones sacadas de los mismos, en el próximo capítulo. En la **Tabla 5** se indican cada uno de estos sistemas.

	Dependientes del locutor	**Independientes del locutor**
A partir de imágenes	Sistema parcial 1	Sistema parcial 3
A partir de voz.	Sistema parcial 2	Sistema parcial 4
A partir de imagen y voz	Sistema parcial 5	Sistema definitivo

Tabla 5. Resumen de los diferentes sistemas implementados.

Como podemos comprobar en dicha tabla existen dos formas de clasificar estos sistemas:

- Según la naturaleza de los datos empleados (imagen, audio o video), los sistemas se dividen en sistemas de reconocimiento de emociones a partir de imágenes, a partir de voz o a partir de imagen y voz.
- Según la dependencia del locutor, los sistemas se dividen en sistemas de reconocimiento de emociones dependientes del locutor o independientes del locutor. En los primeros, la fase de entrenamiento y la de prueba se realizan con observaciones del mismo locutor, mientras que en los segundos, las observaciones de la fase de prueba son de un locutor distinto al de las observaciones de entrenamiento. Esto hace pensar que, a priori, un sistema dependiente del locutor tendrá mejores prestaciones que uno independiente.

El sistema definitivo es un sistema de reconocimiento de emociones independiente del locutor a partir de imagen y voz.

3.1 Esquema General.

Basándonos en lo explicado en el apartado 2.4.1, en la Figura 5 se define el esquema general del sistema de reconocimiento de emociones.

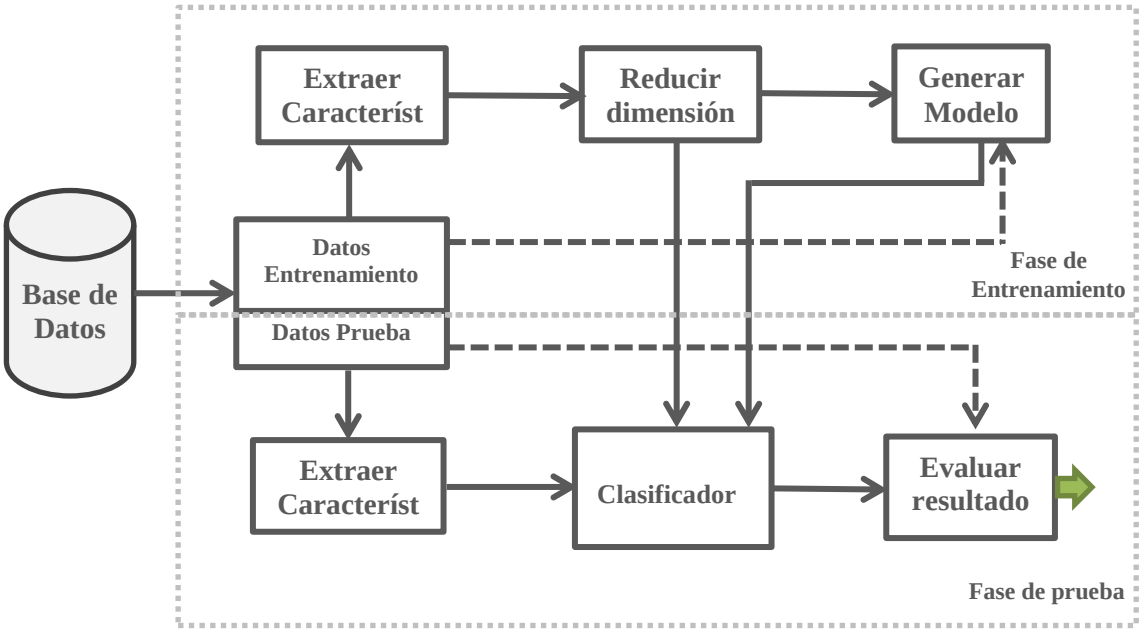


Figura 5. Diagrama de bloques general del sistema de reconocimiento de emociones. Las flechas punteadas simbolizan las etiquetas.

Este esquema general puede adaptarse para los diferentes sistemas parciales antes mencionados modificando la implementación de alguno de sus bloques. Por ejemplo, la diferencia entre los sistemas basados en imágenes y los sistemas basados en audio, se hallará principalmente en el bloque de extracción de características. Por otro lado, la

diferencia entre los sistemas dependientes del locutor y los independientes del locutor yacerá en el bloque que divide los datos en datos de entrenamiento y datos de prueba.

A lo largo de este capítulo se irán explicando detenidamente cada uno de los bloques, indicando que reciben en su entrada y que obtienen a la salida.

3.2 Base de datos.

La base de datos elegida para este trabajo es la *Surrey Audio-Visual Expressed Emotion (SAVEE) Database* [11] creada por los investigadores de la universidad de Surrey, Philip Jackson and Sanaul Haq, como prerrequisito de su sistema de reconocimiento de emociones.

Esta base de datos ha sido creada grabando en audio y vídeo a 4 hombres nativos ingleses (identificados a partir de ahora como DC, JE, JK, KL) con edades comprendidas entre los 27 y 31 años, en un laboratorio de visión 3D. Las emociones se han representado en categorías discretas basadas en las 6 emociones básicas de Ekman comentadas en el apartado 2.1.1 (enfado, asco, miedo, felicidad, tristeza y sorpresa), a las cuales se les ha añadido una nueva categoría para representar la emoción neutral.

Para la creación de la base de datos se les pidió a los 4 locutores que recitasen frente a la cámara y el micrófono hasta un total de 15 frases para cada una de las 7 categorías: 3 frase comunes para todas las emociones, 2 frases específicas de cada emoción y 10 frases genéricas que eran diferentes para cada emoción. Además, las 12 frases específicas de cada emoción (2 por cada una de las 6 emociones) y las 3 comunes fueron grabadas en modo neutral, para obtener un total de 30 frases neutrales. El resultado final es un conjunto de 120 frases (6 x 15 + 30) por cada locutor, lo que significa que la base de datos está formada por un total de 480 frases grabadas en audio y vídeo.

Una de las ventajas de la utilización de esta base de datos es que para la extracción de características de la expresión facial, la cara de los actores fue pintada con 60 marcas azules en ciertos puntos clave de la frente, las cejas, los labios o la mandíbula, como se puede ver en la Figura 6. Gracias a esto y mediante posprocesado de la señal las coordenadas 2D de estas marcas fueron extraídas para cada *frame*, siendo la tasa de refresco de los vídeos de 60 fps (*frames* por segundo). Esto simplifica el trabajo, ya que en vez de trabajar con pixeles de imágenes, únicamente trabajamos con las coordenadas de estas marcas (60 marcas x 2 dimensiones = 120 variables por cada *frame*), que nos dan información sobre el movimiento de diferentes músculos de la cara.

Respecto al audio, la base de datos dispone de los archivos de audio de las 480 frases sin ningún tipo de compresión y con una frecuencia de muestreo de 44, 1 KHz.

Figura 6. Ejemplo de marcas azules en la cara de los locutores con diferentes emociones. De izquierda a derecha: KL (enfadado), JK (feliz), JE (triste) y DC (neutral) [11]

3.3 Partición de las instancias

Este primer bloque se encarga de dividir el conjunto de observaciones procedente de la base de datos en dos conjuntos diferentes: el conjunto de entrenamiento y el conjunto de prueba.

Como ya hemos comentado, la implementación de este bloque difiere si el sistema es dependiente del locutor o independiente del locutor.

3.3.1 Sistemas dependientes del locutor.

En este tipo de sistemas, en realidad, lo que se implementa son 4 sistemas de reconocimiento diferentes para cada uno de los locutores. Cada uno de estos sistemas cuenta con 120 instancias para ser entrenado y probado. Para dividir este conjunto de instancias para cada uno de los locutores se ha empleado lo que se conoce como una validación cruzada de tipo *K-fold*.

La validación cruzada es una técnica utilizada para evaluar los resultados de problemas de aprendizaje supervisado que garantiza que éstos son independientes de la partición de las instancias. En el caso de la validación cruzada de tipo *K-fold*, lo que se hace es dividir el conjunto de instancias en K subconjuntos iguales, de manera que 1 subconjunto se utiliza como conjunto de prueba y los K -1 subconjuntos restantes se utilizan como conjunto de entrenamiento. Este proceso se repite K veces utilizando como conjunto de prueba un subconjunto diferente en cada repetición. El resultado será la media aritmética de los resultados parciales de cada repetición. En la Figura 7 puede verse lo comentado de una manera gráfica.

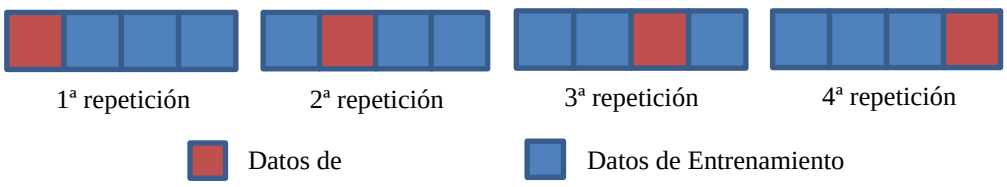


Figura 7. Ejemplo de partición de las instancias mediante validación cruzada 4-fold

En nuestro sistema se ha elegido una K=4, de manera que por cada repetición el conjunto de entrenamiento está formado por 90 instancias y el conjunto de test por 30. En la Figura 8 se muestra un esquema de este bloque del sistema, indicando las dimensiones de las entradas y las salidas.

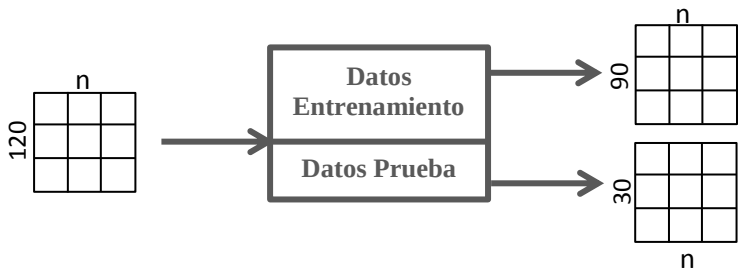


Figura 8. Partición de las instancias en sistemas dependientes del locutor donde n representa el tamaño de las instancias.

3.3.2 Sistemas independientes del locutor.

En este tipo de sistemas la partición es más sencilla. Simplemente utilizamos como conjunto de prueba todas las instancias de un determinado locutor, en nuestro caso 120 instancias, y como conjunto de entrenamiento las de los 3 locutores restantes, en nuestro caso 360 instancias. Este proceso se repite 4 veces utilizando como conjunto de prueba las instancias de un locutor diferente en cada repetición. El resultado final se obtiene haciendo la media aritmética de los resultados parciales de cada repetición. Este proceso se conoce como Leave One Speaker Out (LOSO), que en español significa algo así como "dejar un locutor fuera". En la Figura 9 se muestra el esquema para esta caso.

Esto tiene que ser así para garantizar la independencia del locutor del sistema, ya que las prestaciones del sistema son evaluadas con frases de un locutor cuyas frases no están en el conjunto de entrenamiento del sistema.

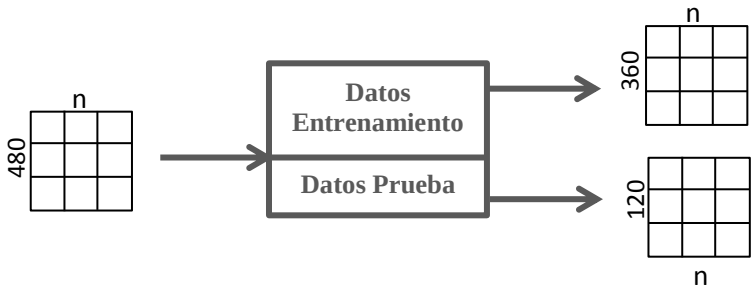


Figura 9. Partición de las instancias en sistemas independientes del locutor donde n representa el tamaño de las instancias.

3.4 Extracción de características.

Este bloque se encarga de extraer características relevantes para el reconocimiento de emociones de los datos proporcionados por la base de datos. Como ya hemos comentado, según la naturaleza de estos datos (audio o vídeo) la extracción de características se lleva a cabo de manera diferente.

3.4.1 Sistemas basados en imágenes.

Como hemos comentado en el apartado 3.2, en el caso de los sistemas basados en imágenes cada observación procedente de la base de datos consta de los valores de las coordenadas en dos dimensiones de las 60 marcas pintadas en las caras de los 4 locutores por cada *frame* de vídeo de la correspondiente frase. Esto hace un total de 120 variables para cada *frame*. Dependiendo de la duración del vídeo, cada frase tendrá diferente número de variables.

Para la extracción de características en este tipo de sistemas se ha decidido hacer la media (véase Ecuación 1) y la desviación típica (véase Ecuación 2) de los valores de las 120 variables en cada uno de los *frames* pertenecientes a una frase. De esta manera se obtienen 120 medias y 120 varianzas, haciendo un total de 240 características por cada frase.

$$\mu[i] = \frac{1}{N}\sum_{n=0}^{N-1} x_i\,[n] \qquad para\ i = 1,2,3 \dots 120$$

Ecuación 1. Cálculo de la media.

$$\sigma[i] = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}(x_i[n] - \mu[i])^2} \qquad para\ i = 1,2,3 \dots 120$$

Ecuación 2. Cálculo de la desviación típica.

donde:

- μ es el vector de medias. Tamaño: 1x120
- σ es el vector de desviaciones típicas. Tamaño: 1x120
- x_i es el vector que representa los valores de la variable i para cada *frame* del vídeo. Tamaño: 1xN
- N es el número de *frames* del vídeo.

De esta manera el vector de características para cada una de las instancias vendría dado por la unión del vector μ y el vector σ, dando como resultado un vector de longitud 240. En la Figura 10 se puede ver un esquema de este bloque.

Figura 10. Extracción de características en sistemas basados en imágenes donde m es el número de observaciones y dependerá de la partición de instancias

3.4.2 Sistemas basados en voz.

En el apartado 2.2 se nombraron varias características de la voz que pueden ser medidas y dar información valiosa sobre el estado emocional del locutor. De todas ellas, en este trabajo nos hemos centrado en trabajar con los Coeficientes Mel-Cepstrales (Mel Frequency Cepstral Coefficients, MFCCs) y con la frecuencia fundamental.

Coeficientes Mel-Cepstrales. [29], [30]

Los MFCCs son uno de los tipos de características más utilizados en el reconocimiento automático del habla (*Automatic Speech Recognition*, ASR). En los últimos años son numerosas las investigaciones que tratan de evaluar su comportamiento para el reconocimiento de emociones.

En la Figura 11, se representa el diagrama de bloques básico para la extracción de estos coeficientes, y a continuación se explicarán brevemente cada uno de los bloques.

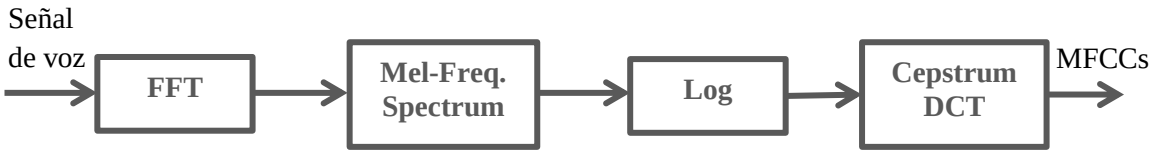


Figura 11. Diagrama de bloques para la extracción de los MFCCs.

El primer bloque se encarga de hacer la Transformada Discreta de Fourier (*Discrete Fourier Transform*, DFT), para obtener la representación en el dominio de la frecuencia de la señal. En la implementación final, lo que en realidad se lleva a cabo es un algoritmo conocido como FFT (*Fast Fourier Transform*), que implementa la Transformada Discreta de Fourier de manera más rápida y eficiente, y el calculo de su módulo. Para llevar a cabo esta transformada se utiliza un enventanado de 20 ms de duración con solapamiento de 10 ms. Esto significa que la señal se ha dividido previamente en segmentos de 20 ms de manera que los segmentos adyacentes esten solapados 10 ms.Los MFCCs han sido calculados para cada uno de estos segmentos. En la Figura 12 puede verse un ejemplo de como se lleva a cabo este enventanado.

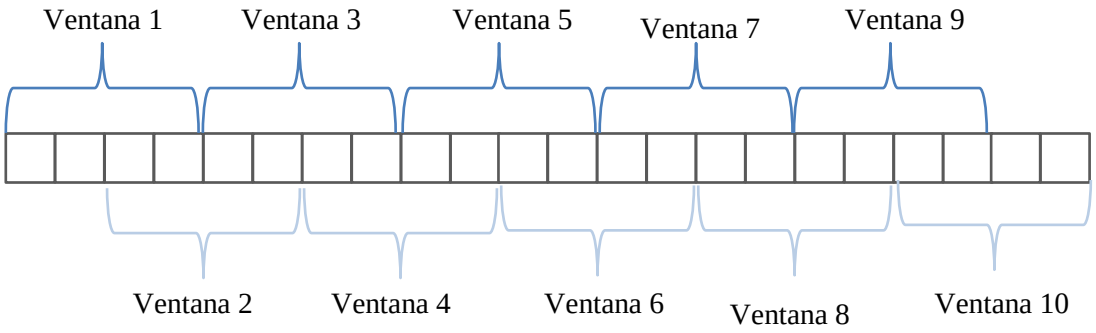


Figura 12. Enventanado con tamaño de ventana de 20 ms y solapamiento de 10 ms.

El segundo bloque es un banco de filtros triangulares en la escala de Mel. Se utiliza esta escala porque trata de imitar la manera no lineal que tiene el oído humano de percibir los sonidos, con una mayor resolución en bajas frecuencias que en altas frecuencias. El resultado de este filtrado es obtener la energía acústica presente en cada una de las bandas de frecuencias del banco de filtros. En la **Figura 13** se muestra un ejemplo de este tipo de filtros.

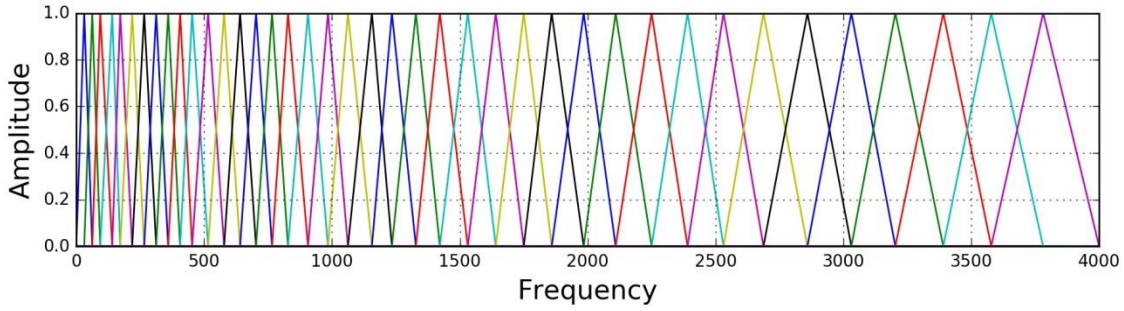


Figura 13. Banco de filtros triangulares en la escala de Mel [30].

El siguiente bloque simplemente calcula el logaritmo de los valores obtenidos en el bloque anterior. Esto se debe a que los humanos percibimos el volumen en una escala logarítmica.

El último bloque trata de eliminar la depencia de la frecuencia fundamental del locutor de la señal. Para ello, calcula el cepstrum del espectro de la señal. El cepstrum de una señal es la Transformada de Fourier Inversa del logaritmo de la Transformada de Fourier de dicha señal (Ver Ecuación 3). Sin embargo, en vez de utilizar la Transformada de Fourier se suele utilizar la Transformada Discreta del Coseno (*Discrete Fourier Transform*, DCT), debido a que en el cálculo de los MFCC sólo se ha considerado el módulo del espectro y por tanto, sus valores son reales y simétricos. La idea es que al hacer el espectro del espectro, la frecuencia fundamental del locutor y sus armónicos se transformen en coeficientes cepstrales de orden mayor que los del resto de la señal. De esta manera, si se realiza un filtro paso bajo después de realizar la transformada inversa,

es posible suprimir estos armónicos y obtener un espectro de la señal libre de toda dependencia con el locutor. Normalmente se eligen como MFCCs los coeficientes entre el 1 y el 13, que hacen un total de 12 MFCCs.

$$F^{-1}\{\log(F\{f(x)\}\}$$

Ecuación 3. Cepstrum de una función f(s).

Este tipo de coeficientes no ofrecen información sobre la variación en el tiempo de la distribución frecuencial de las señales estudiadas. En algunos casos, como en el reconocimiento de emociones es de gran utilidad este tipo de información. Para ello se pueden obtener las derivadas, de primer o de segundo orden, de los coeficientes, lo que nos dará información sobre la velocidad o la aceleración con la que varían en el tiempo los coeficientes.

Para nuestro sistema de reconocimiento, se realizó la media y la desviación típica de las siguientes características basadas en los MFCCs, para cada una de las frases:

- Los 12 MFCCs resultado de realizar el análisis citado anteriormente (24 características).
- La log-energía de cada trama (2 características).
- Los 12 coeficientes resultantes de realizar la primera derivada (24 características).
- La primera derivada de la log-energía (2 características).

En total, 52 características basadas en el estudio de los Coeficientes Mel-cepstrales fueron extraídas para cada una de las observaciones.

Frecuencia fundamental

Si se representa una señal de voz en el dominio del tiempo, se puede comprobar que en los sonidos tonales o sonoros (vocales, semivocales, nasales) se repite un cierto patrón de manera periódica (Ver Figura 14). La frecuencia con la que aparecen estos patrones es lo que se conoce como frecuencia fundamental.

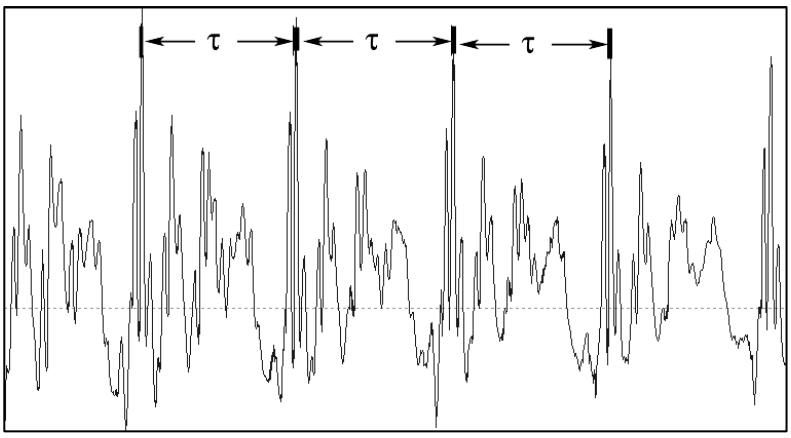


Figura 14. Representación en el dominio del tiempo de una señal de voz. Existe un patrón que se repite cada τ o 1/f segundos [34].

Esta frecuencia depende del aparato fonador de cada locutor, en concreto del movimiento de la glotis, que actúa como una especie de modulador de las vibraciones procedentes de las cuerdas vocales. Esto hace que sea una medida con gran dependencia del locutor y que sea muy útil a la hora de reconocerlo. Sin embargo, numerosos estudios como el llevado a cabo por Carlos Busso en 2009 [31], han demostrado que existe una gran correlación entre la frecuencia fundamental y ciertos estados emocionales.

Para nuestro sistema se han extraído la media y la desviación típica de la frecuencia fundamental para cada una de las frases, en total 2 características por frase.

En la Figura 15 podemos ver cómo queda el bloque de extracción de características para sistemas basados en voz.

Figura 15. Extracción de características en sistemas basados en voz.

3.4.3 Sistemas basados en imagen y voz.

Como se comentó en el apartado 2.4.3 existen diferentes formas de fusionar la información visual y auditiva. En este caso se ha optado por hacer una fusión a nivel de características.

De esta manera las características faciales fueron extraídas conforme al apartados 3.4.1 y las características vocales conforme al apartado 3.4.2, para más tarde ser agrupadas en una única matriz de vectores de características. La Figura 16 muestra el esquema para este tipo de sistemas.

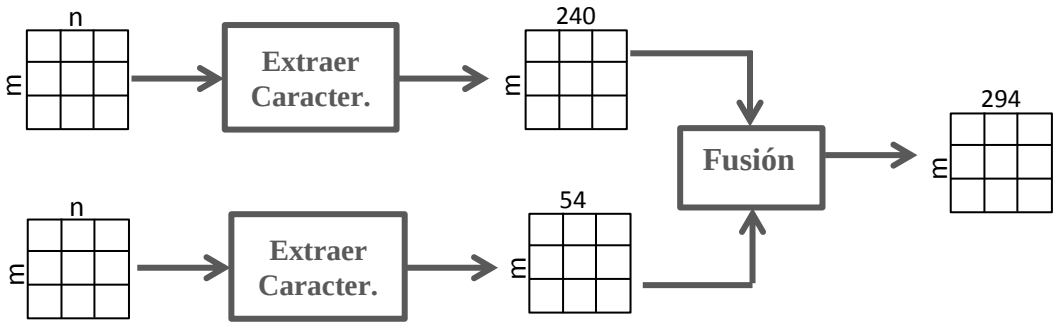


Figura 16. Extracción y fusión de características en sistemas basados en imagen y voz.

Un detalle muy importante a tener en cuenta, que trajo problemas en el proceso de implementación, es la normalización de las características después de ser agrupadas, ya que, al pertenecer a datos de naturaleza distinta, sus rangos pueden ser totalmente distintos lo que genera problemas en algunos clasificadores.

3.5 Reducción de la dimensión [32].

Aunque en el apartado 2.4.1 no fue explicado, la reducción de la dimensión es un bloque típico de cualquier sistema de aprendizaje supervisado. Este proceso consiste en transformar las instancias a un espacio de menor dimensión sin que las prestaciones del sistema se vean mermadas. De manera simplificada, esto es posible debido a que existen ciertas características que son más útiles para la resolución del proceso de aprendizaje que otras, pudiendo incluso haber algunas características que den información redundante o que tengan una alta correlación con otras. Si desechamos estas características redundantes y nos quedamos únicamente con las más útiles, la dimensión del espacio se verá reducida y las prestaciones no se verán afectadas.

La técnica para reducir la dimensión que hemos utilizado para nuestro sistema, y una de las más utilizadas en general, es el Análisis de Componentes Principales (*Principal Component Analysis*, PCA). El PCA es un proceso estadístico que, como su propio nombre indica, se basa en encontrar las componentes principales de un conjunto de muestras. Las componentes principales de un conjunto de muestras son las direcciones del espacio en las cuales la varianza o dispersión de las muestras es máxima. Para entenderlo de manera sencilla, si pensamos en un conjunto de muestras representadas en un espacio de dos dimensiones, la dirección que representa la componente principal de dicho conjunto será la de la recta en la que las proyecciones de los puntos estén más dispersas. En la Figura 17 aparecen dos rectas distintas sobre las que se proyecta el mismo conjunto de puntos. En este ejemplo sencillo, es claro que la dirección de la horizontal es la componente principal.

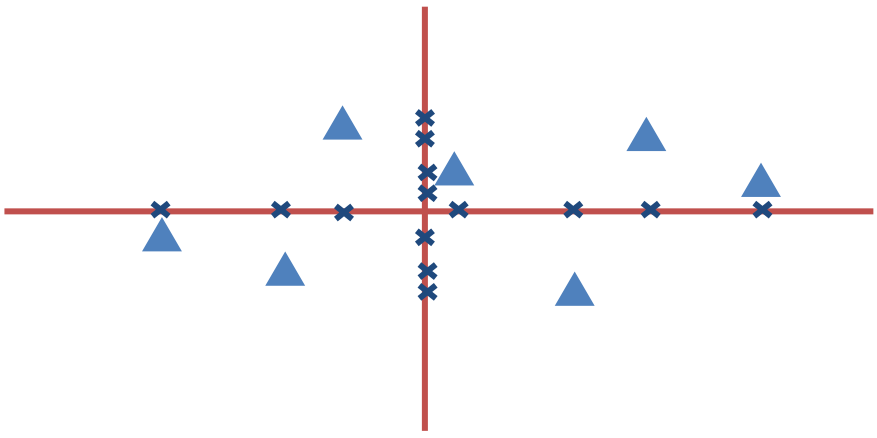


Figura 17. Conjunto de puntos proyectado sobre dos rectas. Las proyecciones sobre la recta horizontal están más dispersas que las proyecciones sobre la recta vertical.

Lo que realmente hace el PCA es una serie de cálculos matemáticas sobre los puntos que recibe a la entrada para obtener una serie de autovectores y autovalores a la salida. Estos autovectores y autovalores vienen dados en pares (el primer autovalor corresponde con el primer autovector y así sucesivamente), y se generan tantos pares como dimensiones tienen las muestras a la entrada. Los autovectores representan direcciones en el espacio y los autovalores dan una medida de la información que proporciona su correspondiente autovector a la hora de representar el conjunto de puntos. De manera que los autovectores con autovalores más altos serán las componentes principales y los autovectores con autovalores cercanos a cero representarán autovectores con información redundante o algún tipo de correlación o dependencia lineal con otros autovectores.

¿Y cómo utilizar el PCA para reducir la dimensión de nuestras observaciones? La idea fundamental es trasladar estas observaciones a un nuevo espacio de coordenadas de menor dimensión. Y los ejes de este nuevo espacio de coordenadas van a ser las componentes principales (los autovectores con autovalores más altos) obtenidas mediante el PCA. Para obtener las coordenadas de los puntos en los nuevos ejes bastará con hacer proyecciones mediante el producto escalar. En la Figura 18 se muestra un esquema de este bloque para reducir la dimensión de las características vocales utilizando un PCA con 15 componentes.

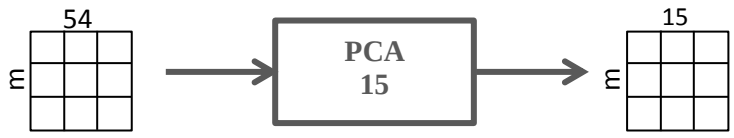


Figura 18. Reducción de la dimensión con PCA 15 en un sistema basado en voz.

3.6 Clasificación

Este es el bloque principal del sistema, que se encarga de generar el modelo o función discriminatoria a partir de las instancias de entrenamiento y sus etiquetas. Como ya vimos en el apartado 2.4.2 existen diferentes algoritmos para llevar a cabo este cometido. De todos ellos, los algoritmos utilizados en este trabajo han sido los siguientes:

- Modelo de Mezclas de Gaussianas (*Gaussian Mixture Model*, GMM)

Para generar el modelo este algoritmo se basa en el modelado de funciones de probabilidad a posteriori que representan la probabilidad de pertenecer a una clase o categoría concreta para cada uno de los posibles valores que pueden tomar cada una de las características. Para modelar estas funciones de probabilidad a posteriori se utiliza la técnica de la mezcla de gaussianas. Esta técnica permite generar la función de probabilidad a posteriori de la primera clase a partir de la nube de puntos que representa

las instancias que pertenecen a dicha clase. Esto se hace gracias a la suposición de que dicha función de probabilidad es una combinación de K distribuciones gaussianas, con lo que el problema se reduce a calcular los parámetros de dichas distribuciones que hagan que se ajusten de la mejor manera posible a la nube de puntos.

Una vez generadas las funciones de probabilidad a posteriori de cada una de las clases o categorías el clasificador lo único que tiene que hacer es, dada una observación, evaluar dicha observación en cada una de las funciones y elegir la categoría cuya función de probabilidad a posteriori evaluada para dicha observación sea mayor.

- Máquina de Vectores Soporte (*Support Vector Machine*, SVM) [33]

Las Máquina de Vectores Soporte o SVM es una de las técnicas más utilizadas para resolver el problema de clasificación. Para entender el objetivo de este algoritmo, lo mejor es ver un ejemplo sencillo en dos dimensiones. Si se considera un conjunto de entrenamiento representado en un espacio de dos dimensiones, de manera que los círculos de color azul pertenecen a una clase y los de color negro a otra, el principal objetivo de la SVM es encontrar una línea que separe perfectamente las observaciones pertenecientes a una clase (círculos azules) de las observaciones pertenecientes a la otra clase (círculos negros) y que esta línea esté lo más alejada posible de los puntos más cercanos a la recta de cada una de las categorías (ver Figura 19). A estos puntos más cercanos a la recta se les denomina vectores soporte y de ahí el nombre de este algoritmo.

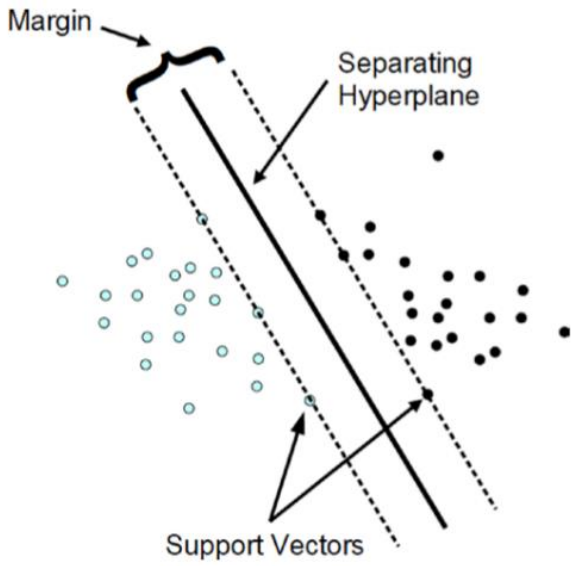


Figura 19. Ejemplo gráfico de la SVM en un espacio bidimensional. Los vectores soporte son los puntos más cercanos al hiperplano [35].

Parece claro que de todas las líneas posibles que separan correctamente las muestras, la línea obtenida por la SVM es la que mejor generaliza, obteniendo la tendencia de los datos y no cayendo en el problema del sobreajuste, comentado en el apartado 2.4.1.

Este ejemplo sencillo sirve para entender el objetivo de la SVM, pero no es un ejemplo demasiado realista. En este ejemplo las muestras únicamente tienen dos dimensiones y se pueden separar perfectamente de manera lineal. En la mayoría de los casos, sin embargo, las muestras son de mayor dimensión no suele ser posible separarlas de manera perfecta e incluso en ocasiones directamente no son linealmente separables. Aun así, la SVM posee técnicas para lidiar con estas situaciones y es aquí donde reside su verdadero potencial.

En caso de que las muestras se encuentren en un espacio de mayor dimensión la única diferencia con nuestro ejemplo es que en vez de buscar líneas rectas, lo que buscará serán hiperplanos que separen correctamente los puntos.

En caso de que las muestras sean separables linealmente pero no de manera perfecta, la SVM define unos márgenes que se ajustan con un parámetro C, de manera que se puede decir al algoritmo qué cantidad de errores en la clasificación estamos dispuestos a permitir. Una vez más, la elección de este parámetro, influirá en lo bien o mal que nuestro modelo generalizará en cómo se comportará con las nuevas muestras de prueba.

En caso de que las muestras no sean linealmente separables, la SVM dispone de una herramienta muy potente: los *kernels*. La idea detrás de esta herramienta es que cuanto mayor sea la dimensión del espacio en el que se encuentran las muestras, más linealmente separables serán. Los *kernels* son una manera eficiente que tiene la SVM para transformar los datos a un espacio de mayores dimensiones y encontrar en este nuevo espacio un hiperplano que los divida linealmente (ver Figura 20)

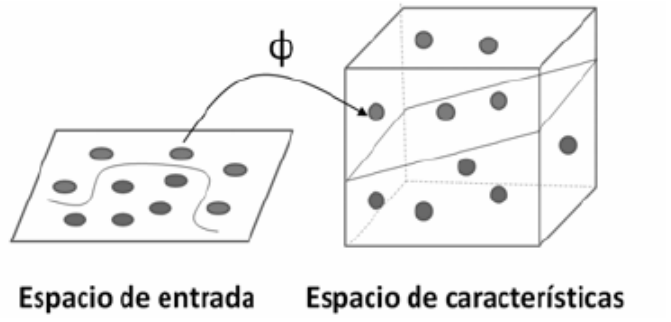


Figura 20. Ejemplo de transformación realizada por un kernel. Los puntos se hacen linealmente separables al transformarlos a un espacio de mayor dimensión [36].

Hay diferentes *kernels*, para diferentes tipos de transformaciones. En este trabajo hemos elegido el *kernel* de base radial. Este tipo de *kernel* mapea los puntos en un espacio de dimensiones infinitas.

En la Figura 21 se muestra un esquema general del bloque de clasificación.

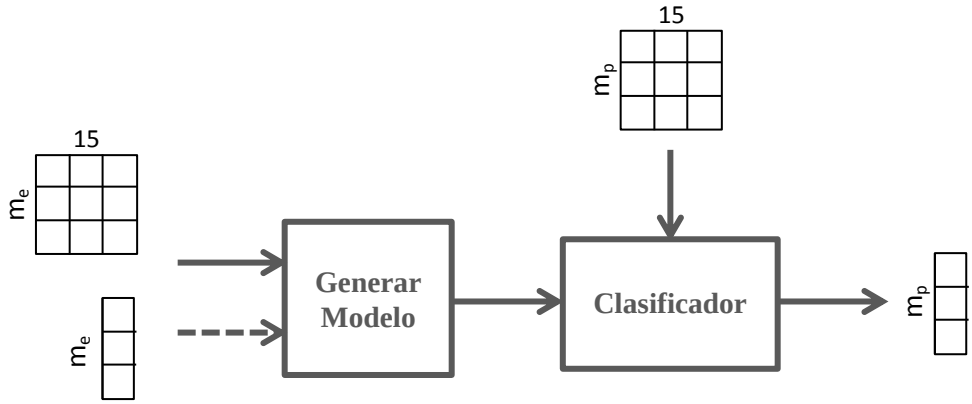


Figura 21. Clasificación. Las observaciones de entrenamiento generan un modelo. El clasificador aplica el modelo sobre las observaciones de prueba y obtiene un vector de etiquetas.

3.7 Evaluación de las prestaciones.

Para evaluar las prestaciones del sistema se utiliza el conjunto de prueba y sus correspondientes etiquetas. Se pasa el conjunto de prueba al clasificador y se compara la salida de éste con el vector de etiquetas del conjunto de prueba.

La medida elegida para medir las prestaciones del sistema ha sido el porcentaje de acierto, que se calcula sumando el número de aciertos (posiciones en las que el vector de salida del clasificador y el vector de etiquetas coincide) y se divide entre el número total de instancias. En la Figura 22 se muestra un esquema de este bloque.

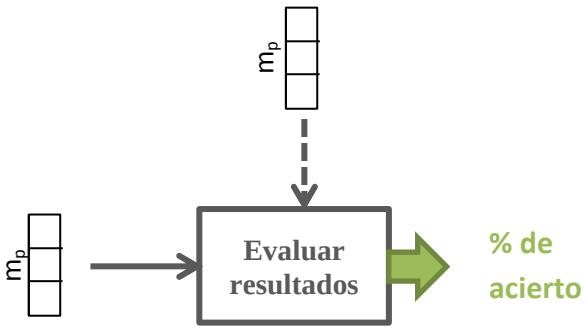


Figura 22. Evaluación de las prestaciones.

Capítulo 4

Experimentos y resultados

En este capítulo se exponen los diferentes experimentos llevados a cabo para evaluar las prestaciones del sistema y se detallan los resultados obtenidos.

Como se ha visto en el Capítulo 3, en el proceso de realización de este trabajo se han implementado varios sistemas parciales antes de la implementación del sistema definitivo. En este capítulo se van a mostrar los resultados obtenidos al evaluar las prestaciones de cada uno de estos sistemas y finalmente se extraerán conclusiones de los mismos. Para facilitar el entendimiento, se utilizarán la misma clasificación y la misma terminología que en el Capítulo 3 (Ver Tabla 5).

Para cada uno de los sistemas se han evaluado las prestaciones de los dos clasificadores vistos en el apartado 3.6 (GMM y SVM), en vistas de descubrir cuál funciona mejor. Además, se han ajustado ciertos parámetros como el número de componentes principales (véase PCA en apartado 3.5) o el número de gaussianas (véase GMM en apartado 3.6).

También cabe mencionar que la forma de llevar a cabo el experimento, y de interpretar los resultados, difiere de si el sistema es dependiente del locutor o independiente del locutor.

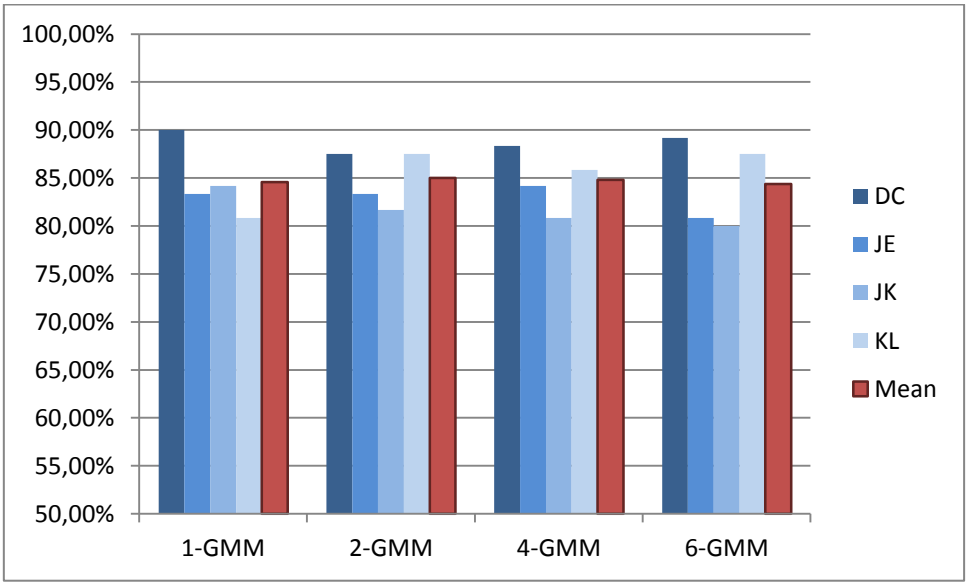
Por cada uno de los sistemas dependientes del locutor, realmente se realizaron 4 experimentos, uno para cada locutor (DC, JE, JK, KL), de manera que por cada uno de estos sistemas se obtienen 4 resultados diferentes, que se refieren al porcentaje de acierto en cada uno de los experimentos. Además se ha realizado la media de estos 4 resultados, para obtener un valor que recoja el comportamiento general del sistema.

Para cada uno de los sistemas independientes del locutor únicamente se realizó un experimento, sin embargo, en cada uno de estos experimentos se evalúan cinco

resultados. Cuatro de ellos, se refieren al porcentaje de acierto del sistema cuando cada uno de los locutores (DC, JE, JK, KL) es elegido como conjunto de prueba, y el último se refiere al porcentaje de acierto total del sistema, que no es más que la media aritmética de los 4 anteriores.

4.1 Sistema parcial 1

El sistema parcial 1 es un sistema de reconocimiento dependiente del locutor y a partir de imágenes. En la Gráfica 1 pueden verse los resultados del clasificador GMM en función del número de gaussianas utilizado, para el caso de utilizar 6 componentes principales (PCA 6). Para este sistema concreto, no pudimos probar valores de PCA mayores debido a que el número de componentes principales no podía ser mayor que el número de observaciones disponibles para generar el modelo (una gaussiana multidimensional por cada emoción). En este caso, el número era muy bajo (entre 5 y 10 observaciones para cada emoción).

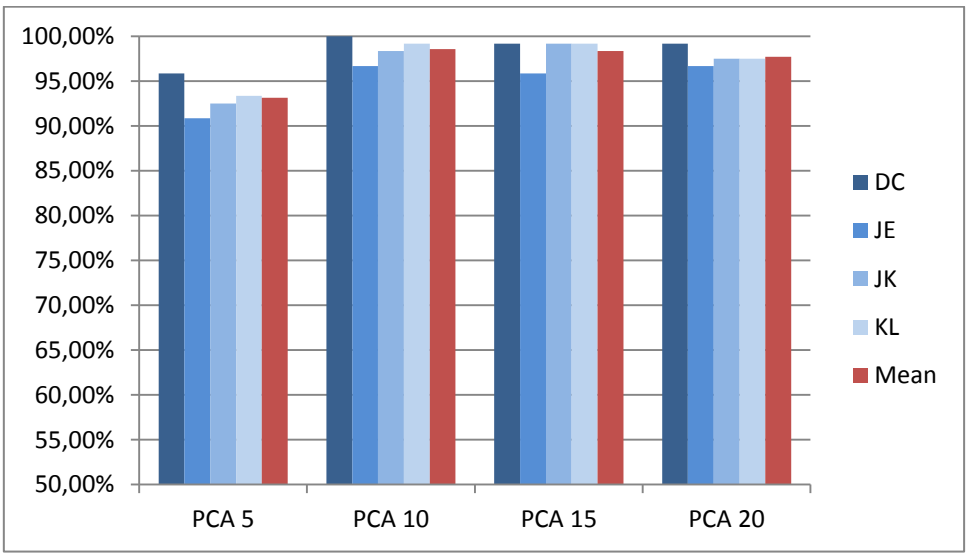


Gráfica 1. Sistema parcial 1. Resultados para GMM con distinto número de gaussianas.

En la Gráfica 2, se han representado los resultados del clasificador SVM en función del número de componentes principales.

Observando la Gráfica 1, vemos que el número de gaussianas elegido no tiene apenas influencia en los resultados. La Gráfica 2 muestra que con PCA 10 el clasificador SVM obtiene los mejores resultados. Comparando ambas gráficas llegamos a la conclusión de que, aunque ambos sistemas ofrecen buenas prestaciones (más del 80% de porcentaje de acierto), el clasificador SVM se comporta considerablemente mejor que el clasificador GMM (98,54% vs 85,00%). Otra observación a destacar es las diferencias existentes en los resultados de las pruebas con diferentes locutores para el mismo sistema. Por lo

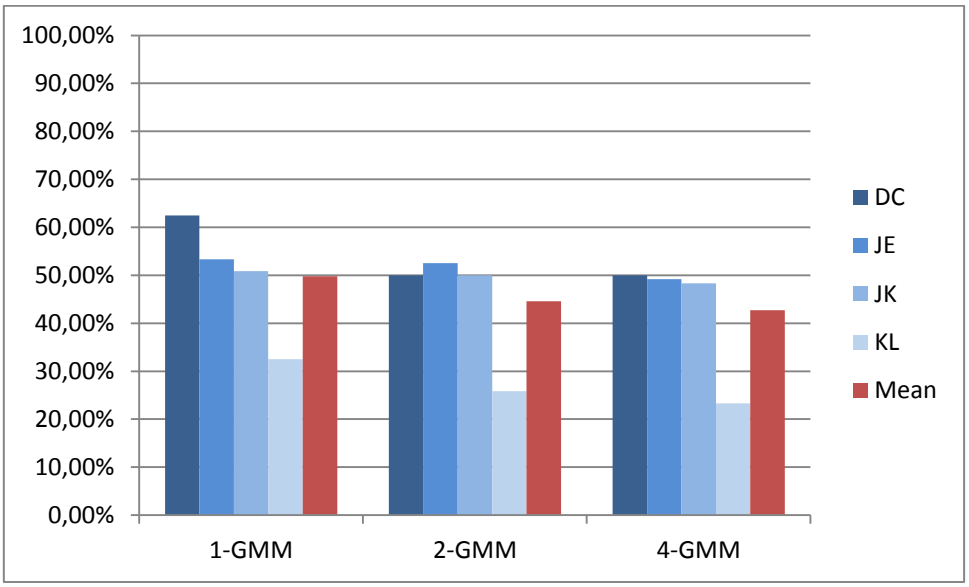
general parece que las pruebas llevadas a cabo con los datos del locutor DC obtienen resultados considerablemente mejores. Esto puede deberse a que dicho locutor sea una persona más expresiva.



Gráfica 2. Sistema parcial 1. Resultados para SVM con distintos PCAs.

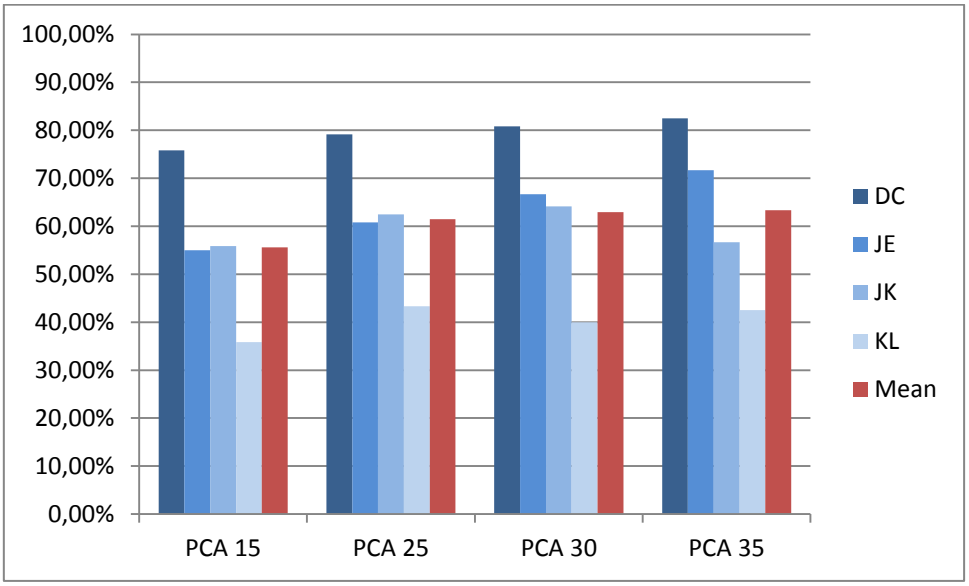
4.2 Sistema parcial 2

El sistema parcial 2 es un sistema de reconocimiento dependiente del locutor a partir de voz. En la Gráfica 3 están plasmados los resultados del clasificador GMM en función del número de gaussianas utilizado, para el caso concreto de utilizar 6 componentes principales (PCA 6). En la Gráfica 4 se han representado los resultados del clasificador SVM para distintos valores de PCA.



Gráfica 3. Sistema parcial 2. Resultados para GMM con distinto número de gaussianas.

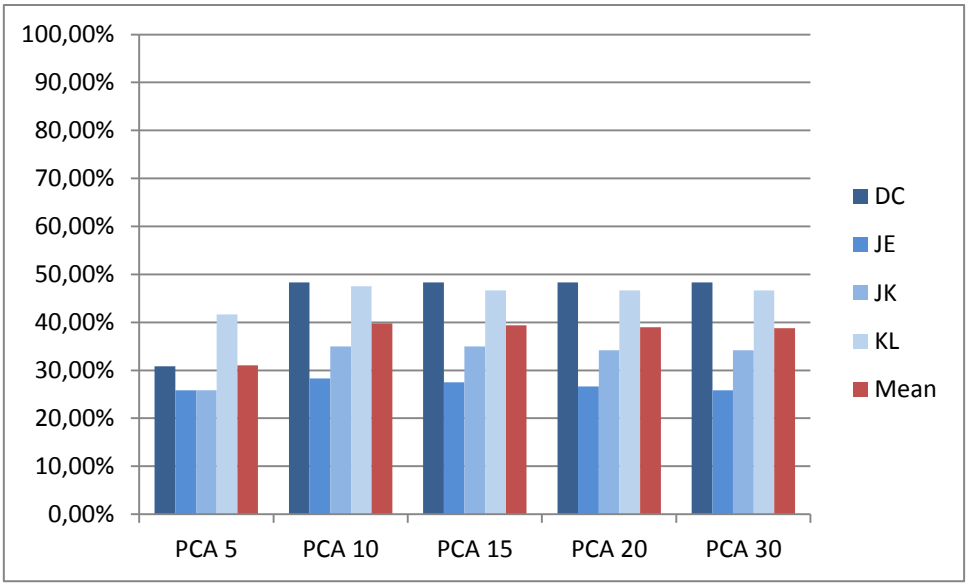
La Gráfica 3 muestra que utilizando el clasificador GMM para el sistema parcial 2, el mejor resultado se obtiene utilizando una única gaussiana (49,79 %) y que los resultados empeoran conforme aumenta el número de gaussianas utilizado. Por otro lado, observando la Gráfica 4 vemos que el clasificador SVM obtiene los mejores resultados para un PCA 35 (63,33%). Comparando ambas gráficas, queda patente que el clasificador SVM ofrece mejores prestaciones que el clasificador GMM (63,33% vs 49,77%). Además, si comparamos la Gráfica 3 y la Gráfica 4 con la Gráfica 1 y la Gráfica 2, vemos que los sistemas basados en imágenes ofrecen mejores prestaciones que los sistemas basados en voz (98,54% vs 63,33% en el caso del clasificador SVM).



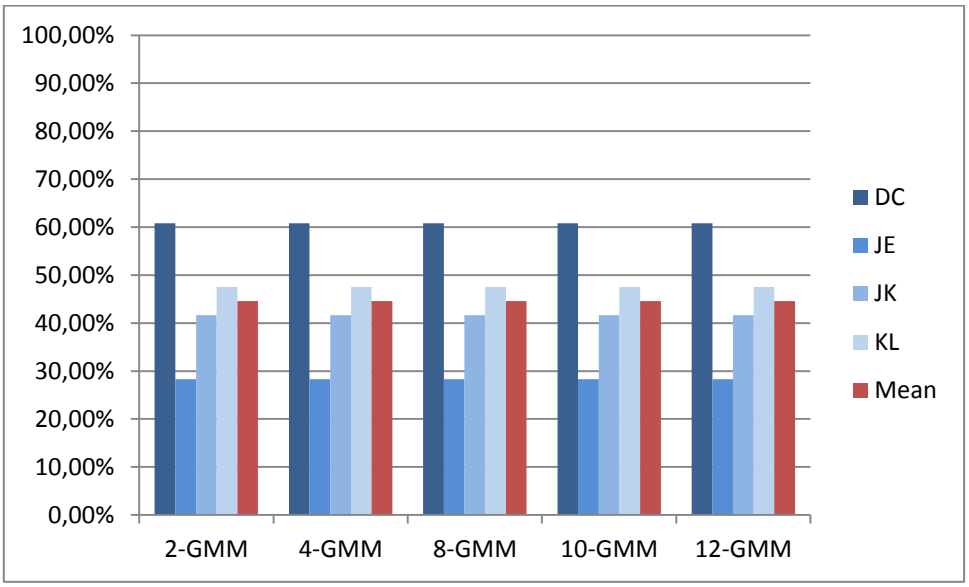
Gráfica 4. Sistema parcial 2. Resultados para SVM con distintos PCAs.

4.3 Sistema parcial 3

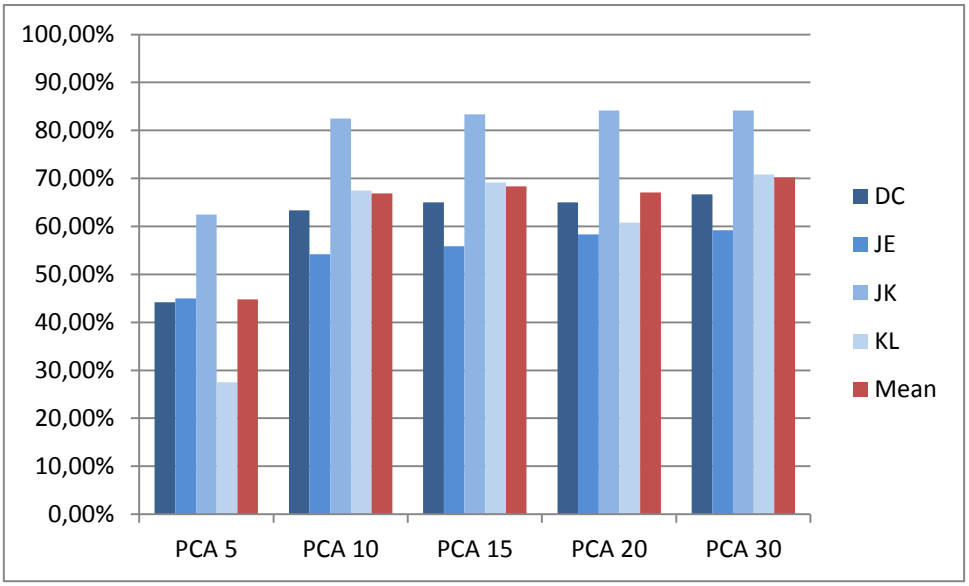
El sistema parcial 3 es un sistema de reconocimiento independiente del locutor a partir de imágenes. En la Gráfica 5 pueden verse los resultados del clasificador GMM con una única gaussiana para diferentes valores de PCA. En la Gráfica 6 se muestran los resultados del clasificador GMM con PCA 10 para diferentes números de gaussianas. En la Gráfica 7 aparecen los resultados del clasificador SVM para diferentes valores de PCA.



Gráfica 5. Sistema parcial 3. Resultados para GMM (1 gaussiana) con distintos PCAs.



Gráfica 6. Sistema parcial 3. Resultados para GMM con distinto número de gaussianas.

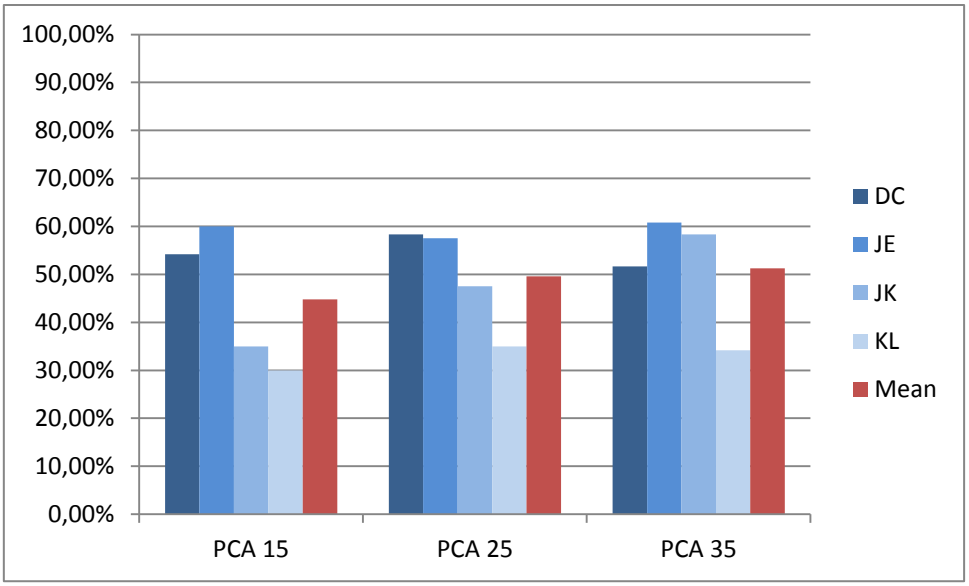


Gráfica 7. Sistema parcial 3. Resultados para SVM con distintos PCAs.

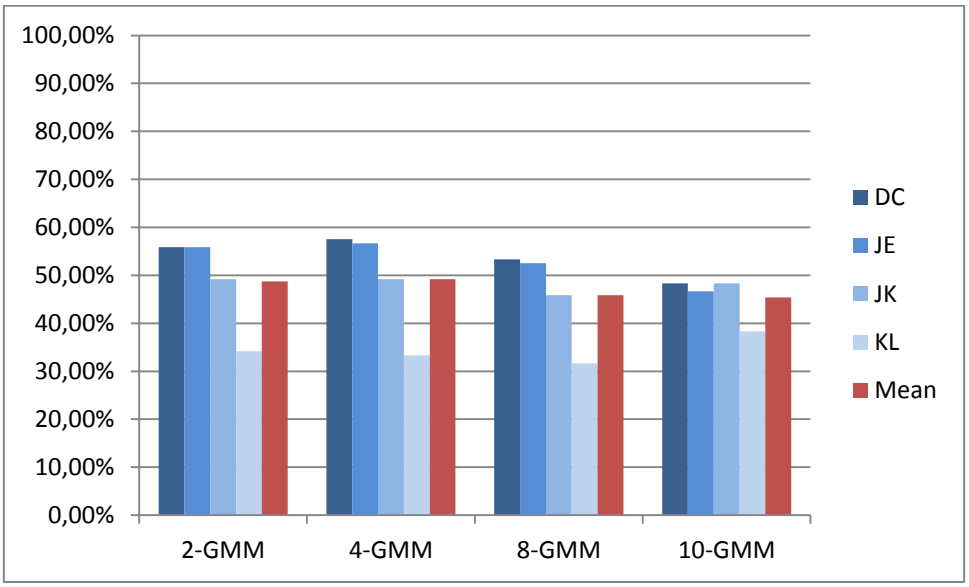
Observando la Gráfica 5, vemos que utilizando el clasificador GMM, para el caso concreto de una única gaussiana, los mejores resultados se obtienen para un PCA 10 (39,78%). La Gráfica 6, por su parte, sugiere que el número de gaussianas utilizado en dicho clasificador no tiene ninguna influencia en el resultado final. Finalmente, en la Gráfica 7 observamos que cuando se utiliza el clasificador SVM los mejores resultados se obtienen para un PCA 30 (70,21%). Comparando la Gráfica 7 con la Gráfica 6, comprobamos que para este tipo de sistemas también funciona mejor el clasificador SVM que el GMM (70,21% vs 39,78%), siendo incluso esta diferencia mayor que en los sistemas parciales 1 y 2. Comparando la Gráfica 7 con la Gráfica 2, se observa como las prestaciones son mucha mayores en sistemas dependientes del locutor que en sistemas independientes del locutor, como era de esperar (98,54% vs 70,21%).

4.4 Sistema parcial 4

El sistema parcial 4 es un sistema de reconocimiento independiente del locutor a partir de voz. En la Gráfica 8 pueden verse los resultados del clasificador GMM con una única gaussiana para diferentes valores de PCA. En la Gráfica 9 se representan los resultados del clasificador GMM con PCA 10 para diferentes números de gaussianas. En la Gráfica 10 aparecen los resultados del clasificador SVM para diferentes valores de PCA.

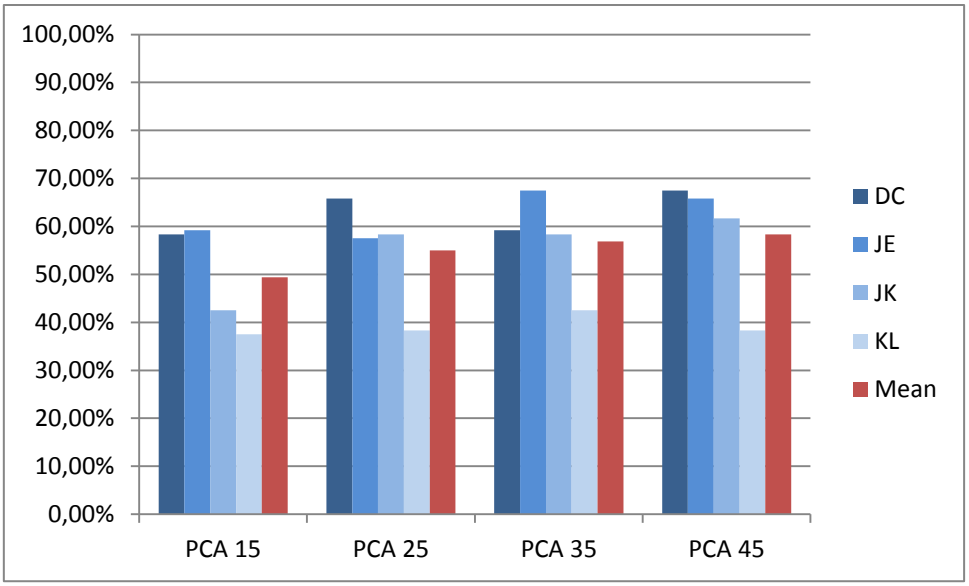


Gráfica 8. Sistema parcial 4. Resultados para GMM (1 gaussiana) con distintos PCAs.



Gráfica 9. Sistema parcial 4. Resultados para GMM con distinto número de gaussianas.

Si observamos la Gráfica 8, vemos que en las pruebas con el clasificador GMM, para el caso concreto de utilizar una única gaussiana, se obtuvo el mejor resultado con un PCA 35 (51,25%). La Gráfica 9 muestra que el aumento del número de gaussianas en el GMM no mejoró el resultado anterior. En la Gráfica 10 se observa que en las pruebas con el clasificador SVM, el mejor resultado se obtuvo con un PCA 45 (58,33%).



Gráfica 10. Sistema parcial 4. Resultados para SVM con distintos PCAs.

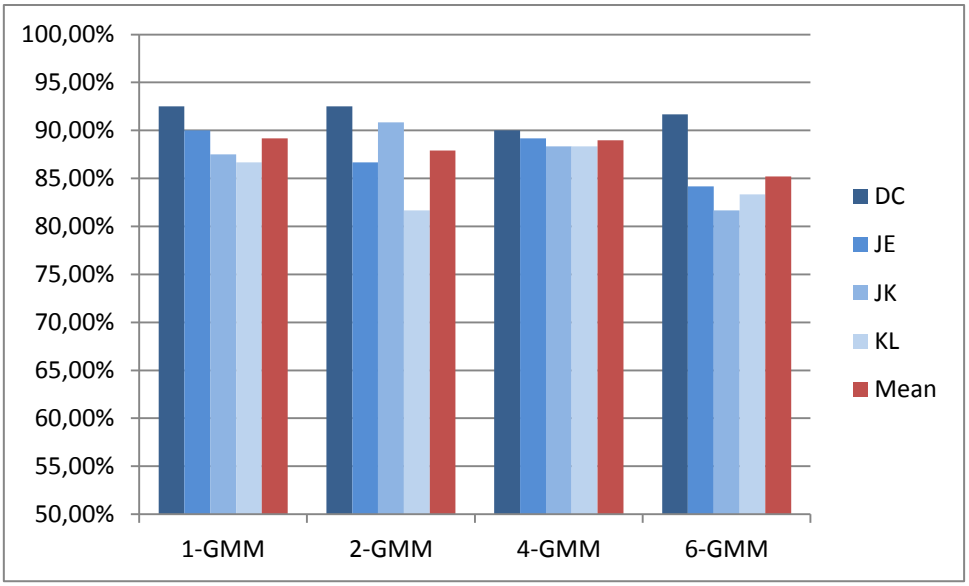
Comparando la Gráfica 10 con las Gráficas Gráfica 8 y Gráfica 9 , comprobamos que, una vez más, el clasificador SVM tiene mejores prestaciones que el clasificador GMM (58,33% vs 51,25%). Si se compara la Gráfica 10 con la Gráfica 7, comprobamos que, también en sistemas independientes del locutor, los sistemas basados en imágenes tienen mayores prestaciones que los sistemas basados en voz (70,21% vs 58,33%). Al comparar la Gráfica 10 con la Gráfica 4, se confirma que los sistemas dependientes del locutor obtienen mejores resultados que los independientes (63,33% vs 58,33%).

4.5 Sistema parcial 5

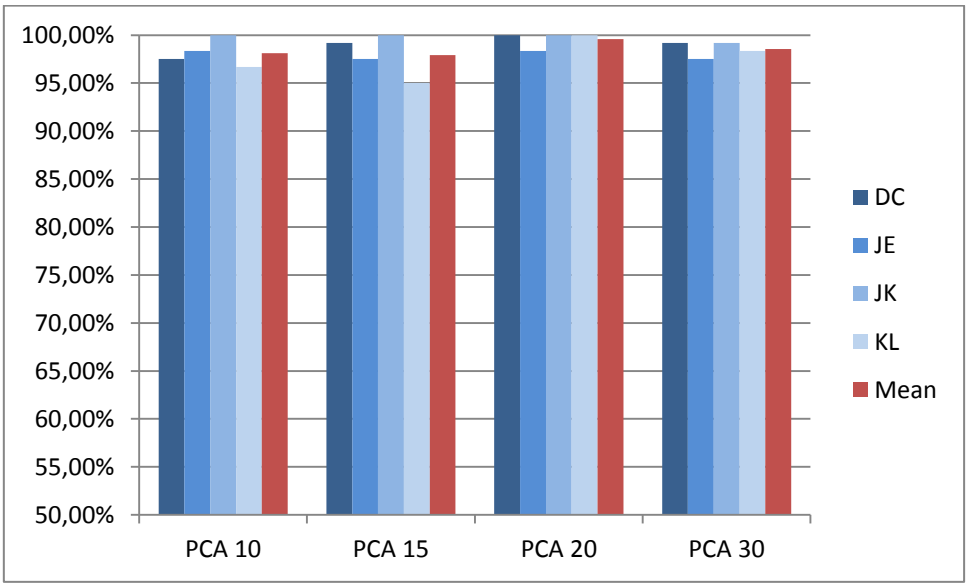
El sistema parcial 5 es un sistema de reconocimiento dependiente del locutor a partir de imagen y voz. En la Gráfica 11 pueden verse los resultados del clasificador GMM con PCA 6 para diferentes números de gaussianas. En la Gráfica 12 aparecen los resultados del clasificador SVM para diferentes valores de PCA.

Si se observa la Gráfica 11, podemos ver que para las pruebas realizadas con el clasificador GMM, para el caso concreto de utilizar un PCA 6, los mejores resultados se obtuvieron utilizando una única gaussiana (89,17%). En la Gráfica 12 puede verse que para las pruebas llevadas a cabo con el clasificador SVM, los mejores resultados se obtuvieron con un PCA 20 (99,58%).

Comparando ambas gráficas, vemos que, aunque ambos tiene un buen comportamiento, el clasificador SVM ofrece mejores prestaciones que el GMM (99,58% vs 89,17%)… Además si se compara la Gráfica 12 con la Gráfica 2 y la Gráfica 4, se observa que la fusión de las características basadas en imagen y voz funciona mejor que cada una de ellas por separado (99,58% vs 98,54% para imagen y 63,33% para voz).



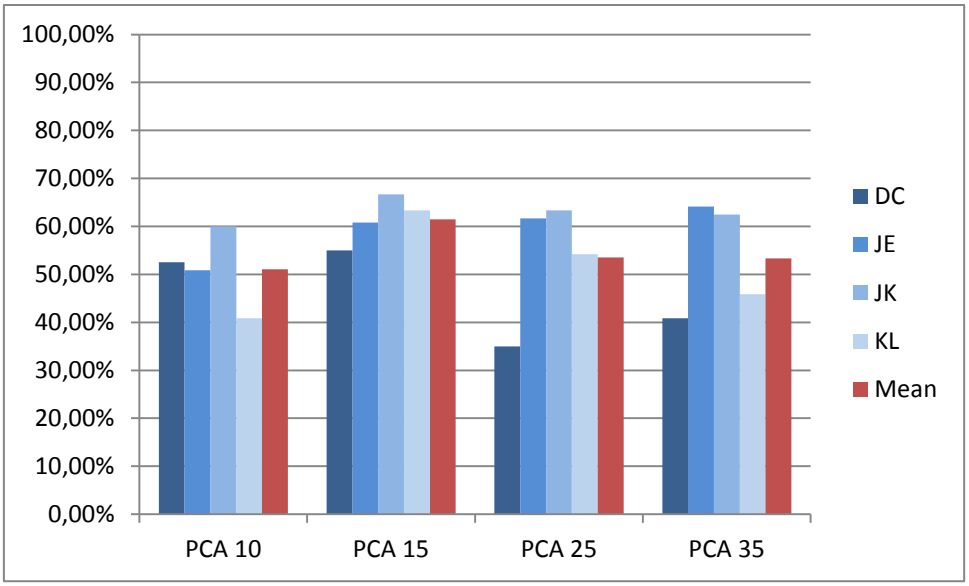
Gráfica 11. Sistema parcial 5. Resultados para GMM con distinto número de gaussianas



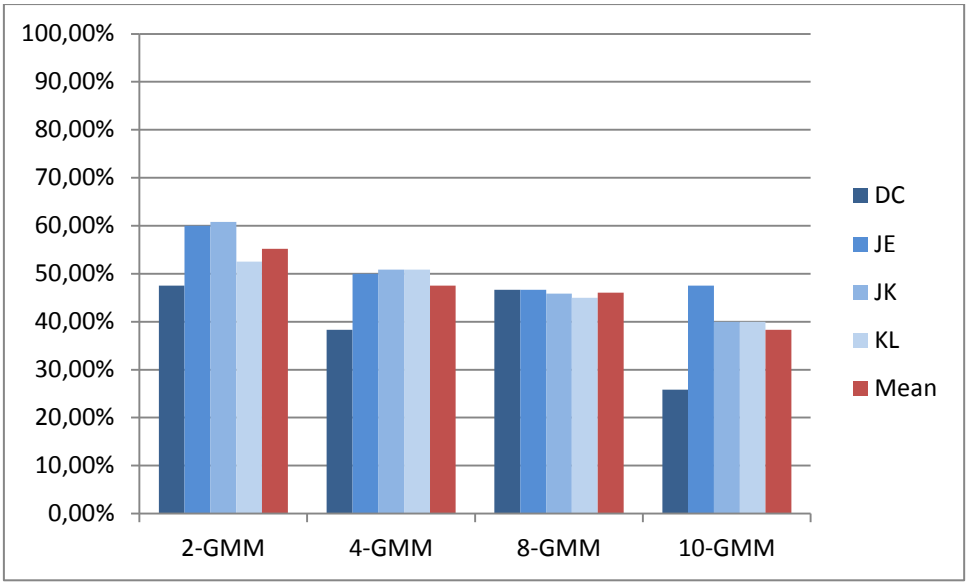
Gráfica 12. Sistema parcial 5. Resultados para SVM con distintos PCAs.

4.6 Sistema definitivo

El sistema definitivo es un sistema de reconocimiento independiente del locutor a partir de imágenes y voz. En la Gráfica 13 pueden verse los resultados del clasificador GMM con una única gaussiana para diferentes valores de PCA. En la Gráfica 14 se muestran los resultados del clasificador GMM con PCA 10 para diferentes números de gaussianas. En la Gráfica 15 aparecen los resultados del clasificador SVM para diferentes valores de PCA.



Gráfica 13. Sistema definitivo. Resultados para GMM (1 gaussiana) con distintos PCAs.

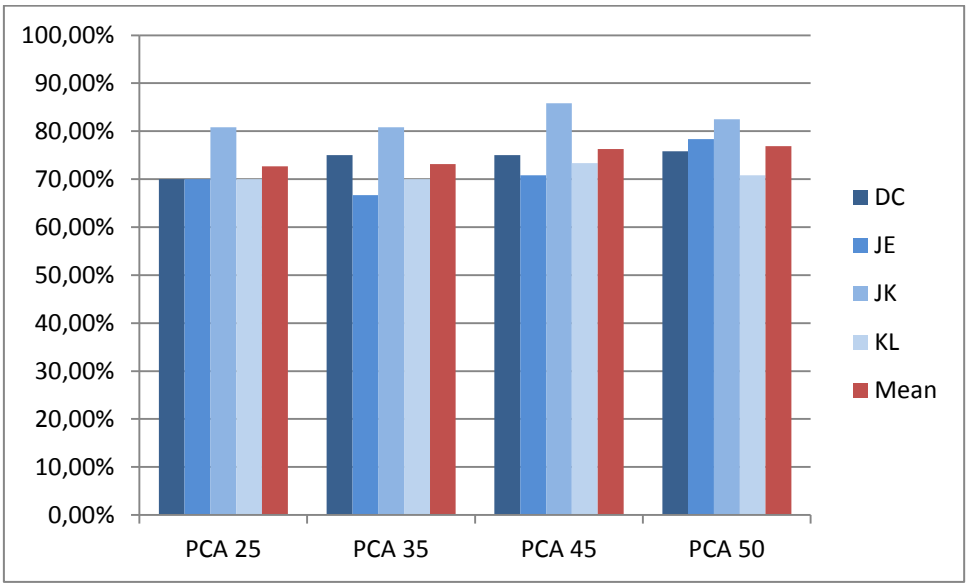


Gráfica 14. Sistema definitivo. Resultados para GMM con distinto número de gaussianas

Si se observa la Gráfica 13 puede verse que en las pruebas realizadas con el clasificador GMM, en el caso concreto de utilizar una única gaussiana, los mejores resultados se obtuvieron con un PCA 15 (61,46%). La Gráfica 14 muestra que en dichas pruebas el aumento de gaussianas utilizadas no mejoró el resultado anterior. En la Gráfica 15 vemos que en las pruebas llevadas a cabo con el clasificador SVM, los mejores resultados fueron obtenidos con PCA 50 (76,88%).

Comparando dichas gráficas, vuelve a ser claro que el clasificador SVM obtiene mejores prestaciones que el GMM (76,88% vs 61,46%). Si se comparan la Gráfica 15 con la Gráfica 7 y la Gráfica 10, se observa que, también en sistemas independientes del

locutor, la fusión de características de audio y vídeo funciona mejor que cada una de ellas por separado, en este caso de forma más notable si cabe (76,88 % para la fusión vs 70,24% para vídeo y 58,33% para audio). Además, al comparar la Gráfica 15 con la Gráfica 12, se confirma también para sistemas a partir de audio y vídeo, que los sistemas dependientes del locutor obtienen mejores resultados que los sistemas independientes del locutor (99,58% vs 76,88%).



Gráfica 15. Sistema definitivo. Resultados para SVM con distintos PCAs.

4.6.1 Matrices de confusión.

En la Tabla 6 están representadas las matrices de confusión del sistema definitivo (con clasificador SVM y PCA 50) para cada uno de los locutores. Una matriz de confusión es una matriz de tamaño MxM, donde M es el número de clases o categorías definidas en la tarea de clasificación, que aloja en la posición (i, j) las veces que una observación etiquetada con la categoría i ha sido clasificada por el sistema como la categoría j, de manera que en la diagonal de la matriz (cuando i = j) aparecen las veces que el sistema ha acertado reconociendo cada una de las categorías. En nuestro sistema de reconocimiento de emociones M = 7 (6 emociones básicas + neutral), de manera que las matrices de confusión son de tamaño 7x7. Para nombrar las distintas categorías se ha utilizado las iniciales de las emociones que representan en inglés: enfado (***A**nger)*, aversión *(**D**isgust)*, miedo *(**F**ear)*, felicidad *(**H**appiness)*, neutra *(**N**eutral)*, Tristeza *(**Sa**dness)*, Sorpresa *(**Su**rprise*). De esta manera, en la posición (A, F) aparecerá el número de veces que el sistema ha clasificado con la categoría *Fear* una observación etiquetada como *Anger*.

	A	D	F	H	N	Sa	Su
A	15	0	0	0	0	0	0
D	7	8	0	0	0	0	0
F	2	0	0	2	0	0	11
H	0	0	0	15	0	0	0
N	0	0	0	0	30	0	0
Sa	0	0	4	0	0	9	2
Su	1	0	0	0	0	0	14

a)

A	D	F	H	N	Sa	Su
15	0	0	0	0	0	0
8	7	0	0	0	0	0
5	0	9	0	0	1	0
0	0	3	12	0	0	0
0	0	0	0	30	0	0
0	0	1	0	5	9	0
0	0	1	0	2	0	12

b)

	A	D	F	H	N	Sa	Su
A	10	0	5	0	0	0	0
D	1	12	0	0	0	2	0
F	0	0	9	1	0	0	5
H	0	0	0	15	0	0	0
N	0	0	0	0	30	0	0
Sa	0	0	1	0	5	9	0
Su	0	0	0	1	0	0	14

c)

A	D	F	H	N	Sa	Su
3	12	0	0	0	0	0
1	12	1	0	0	0	1
0	0	15	0	0	0	0
3	1	0	8	0	0	3
0	0	0	2	28	0	0
0	9	1	0	0	5	0
0	0	0	1	0	0	14

d)

Tabla 6. Matrices de confusión del sistema definitivo para cada locutor: a) DC, b) JK, c) JE, d) KL

Observando las tablas vemos que para el locutor DC, el sistema apenas comete fallos al reconocer emociones como el enfado (A), la felicidad (H), la neutra (N) o la sorpresa (Su) y que por el contrario, la tasa de error es muy elevada para la clase miedo (F). Para el locutor JK vemos que el sistema se comporta de manera aceptable en las mismas emociones que para DC (A, H, N, Su), pero en este caso la emoción que peor se reconoce es la aversión (D). Para el locutor JE, el sistema vuelve a comportarse mejor para las 4 emociones ya mencionadas, además de para la aversión (D), mientras que se comporta ligeramente peor para emociones como el miedo y la tristeza. Finalmente, para el locutor KL el sistema clasifica con mayor acierto las emociones como la aversión (D), el miedo (F), la neutra (N) y la sorpresa (Su), y por el contrario clasifica con menor acierto emociones como el enfado (A) y la tristeza (Sa).

Capítulo 5

Planificación y presupuesto

En este capítulo se explica cuál ha sido la planificación del proyecto, identificando sus diferentes fases y definiendo los respectivos tiempos. Además, se desglosa el presupuesto que ha sido necesario para llevarlo a cabo.

5.1 Planificación.

En la realización de cualquier tipo de proyecto, el proceso de planificación es una fase crucial para definir las diferentes tareas a llevar a cabo, establecer un orden de prioridades e identificar qué márgenes de tiempo existen para la realización de cada una de ellas.

En la Tabla 7 se muestra la planificación estimada para la realización de este proyecto. En ella se indican la fecha de inicio y entrega del proyecto, además de las fechas de inicio de las fases principales (desarrollo del software y escritura de la memoria). También incluye las diferentes tareas a llevar a cabo y su duración estimada en días.

Por su parte, la Figura 23 incluye un pequeño diagrama de Gantt del proyecto, en el que se puede entender de manera gráfica cuál es la estructura del proyecto, la secuencia de las diferentes tareas y la duración relativa de cada una de ellas respecto de la duración total del proyecto.

PLANIFICACIÓN DEL PROYECTO	
Fecha de inicio del proyecto: 26/01/2017	
Tareas	**Duración (días)**
Fecha de inicio del desarrollo software: 26/01/2016	
Documentación (I)	15
Sistema parcial 1	20
Sistema parcial 2	25
Sistema parcial 3	15
Sistema parcial 4	20
Pruebas y resultados (I)	5
Sistema parcial 5	20
Sistema definitivo	15
Pruebas y resultados (II)	5
Fecha de inicio de la escritura de la memoria: 15/05/2017	
Documentación (II)	20
Escritura	55
Vacaciones del 01/08/2017 al 31/08/2017	
Escritura	15
Revisión	10
Fecha de entrega del proyecto: 26/09/2017	

Tabla 7. Planificación del proyecto.

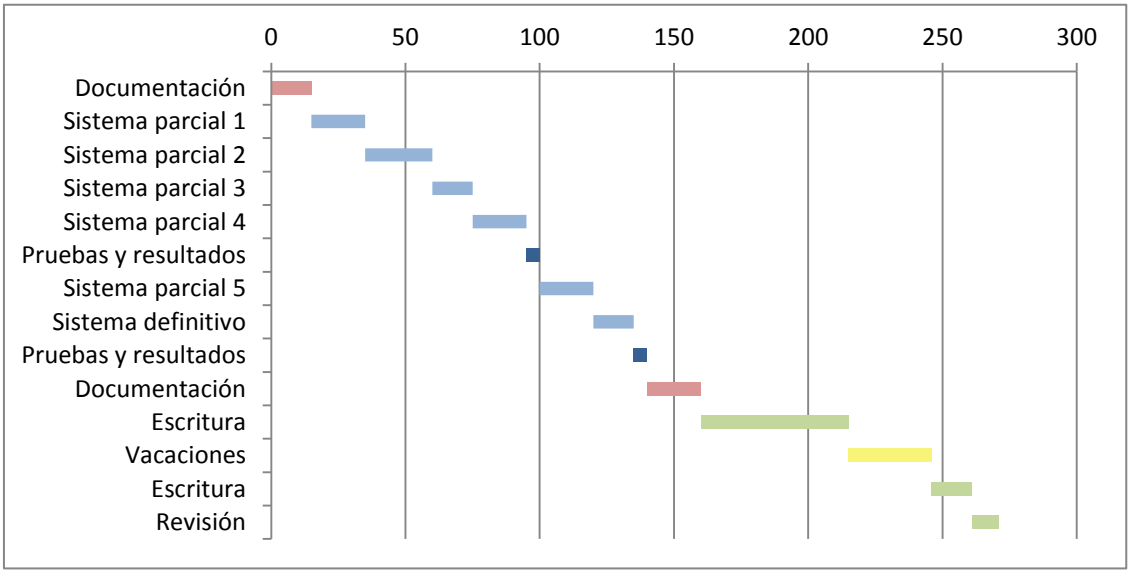


Figura 23. Diagrama de Gantt del proyecto.

5.2 Presupuesto.

A continuación se hace un desglose del presupuesto del proyecto, haciendo una estimación de sus costes. Se han dividido los diferentes costes en: hardware, software y recursos humanos.

- Hardware.

El único hardware utilizado para la realización del proyecto ha sido un portátil Sony Vaio SVE15125CXS, cuyo coste es de 980 €. En la Tabla 8 se indica el coste de amortización de este portátil suponiendo una amortización lineal de 10 años.

Recurso	Coste amortización
Portátil Sony Vaio SVE15125CXS	98 €
TOTAL	98 €

Tabla 8. Presupuesto recursos hardware.

- Software

En la Tabla 9 se recogen los costes de los diferentes recursos software necesarios para la realización del proyecto, incluyendo el sistema operativo.

Recurso	Coste
Microsoft Windows 10 Home 64 Bits	135 €
Matlab R2015a Education License	500 €
TOTAL	635 €

Tabla 9. Presupuesto recursos software.

- Recursos humanos.

Para calcular el coste de los recursos humanos, se ha supuesto que para la realización del proyecto han sido necesarios un desarrollador con nivel de ingeniero junior y un supervisor con nivel de ingeniero senior. En la Tabla 10 vienen desglosadas las horas de trabajo mensuales, el coste asignado por cada hora de trabajo y el coste total para cada

uno de los trabajadores. Se ha tenido en cuenta para la realización de esta tabla que la duración del proyecto ha sido de 8 meses.

Recurso	Horas de trabajo	Coste por hora	Coste total
Desarrollador junior	80 horas/mes	15€/hora	9600
Supervisor	6 horas/mes	20€/hora	960
TOTAL			10560

Tabla 10. Presupuesto recursos humanos.

Finalmente, en la Tabla 11 se muestra el desglose del presupuesto total, cuya cifra asciende a once mil doscientos noventa y tres euros sin IVA.

Recurso	Coste
Recursos hardware	98 €
Recursos software	635 €
Recursos humanos	10560
TOTAL	11.293 €

Tabla 11. Presupuesto del proyecto.

Capítulo 6

Conclusiones y líneas futuras.

En este capítulo se presentan las conclusiones extraídas de la investigación llevada a cabo y se enumeran algunas líneas de investigación futuras que se abren tras la realización del proyecto.

6.1 Conclusiones.

El objetivo principal de este trabajo era la realización de un sistema de reconocimiento de emociones a partir de imagen y voz. Podemos concluir que este objetivo principal ha sido cumplido. Si bien es cierto que, pese a que el sistema de reconocimiento de emociones a partir de imagen y voz dependiente del locutor (sistema parcial 5) ha alcanzado unas prestaciones realmente buenas, en el caso del sistema independiente del locutor (sistema definitivo) las prestaciones obtenidas han sido menores, debido a la dificultad intrínseca de la tarea y a lo reducido de la base de datos.

Los objetivos secundarios en su gran mayoría han sido conseguidos. Para la realización del trabajo se ha llevado a cabo un profundo estudio de las emociones: su naturaleza, su clasificación y su expresión...etc. y de las bases del aprendizaje automático: el proceso de aprendizaje, los diferentes clasificadores, los sistemas de clasificación multimodal...etc. Además, gracias a las diferentes pruebas realizadas han sido evaluadas las prestaciones de diferentes características, tanto vocales como faciales, y de diferentes clasificadores para el problema concreto del reconocimiento de emociones. También han sido evaluadas las prestaciones de un sistema multimodal fusionando las características vocales y faciales.

Tras los numerosos experimentos realizados, a continuación se hace un resumen de las principales conclusiones extraídas:

- El clasificador SVM con kernel RBF ofrece mejores prestaciones que el clasificador GMM, para todos los sistemas evaluados.
- Los sistemas basados en imágenes han ofrecido mejores prestaciones que los basados en voz. Esto quiere decir que las características extraídas a partir de imágenes ofrecen más información que las características extraídas de la voz sobre el estado emocional de la persona.
- Los sistemas basados en imagen y voz han ofrecido mejores prestaciones que los sistemas basados únicamente en imágenes o únicamente en voz. Esto quiere decir que la fusión de los dos tipos de características ofrece más información que cada uno por separado.

6.2 Líneas Futuras.

Debido al tiempo limitado del que se disponía, este trabajo no ha podido seguir todas las líneas de investigación que se iban abriendo durante el proceso. A continuación, se nombran algunas de estas líneas de investigación que podrían ser abordadas en el futuro:

- Para realizar este sistema de reconocimiento de emociones, se ha contado con la ventaja de disponer de las coordenadas de las marcas azules previamente pintadas en las caras de los locutores. Una línea futura podría consistir en extraer estas coordenadas de puntos clave de la cara de una manera automática, sin que sea necesario que los locutores tengan ningún tipo de marca manual.
- Una de las técnicas que está revolucionando el campo del Aprendizaje Automático son las Redes Neuronales Convolucionales (*Convolutional Neural Networks*, CNN). En una línea futura se podría emplear este tipo de redes neuronales, tanto para la extracción de las coordenadas antes mencionadas como para la realización del sistema de reconocimiento. El problema que tienen estas técnicas es que precisan de bases de datos de gran tamaño y, a día de hoy, no existen bases de datos tan grandes para el reconocimiento de emociones. Además, requieren de una gran carga computacional.
- La base de datos que hemos utilizado para implementar nuestro sistema ha sido grabada en condiciones de laboratorio, esto es, con una gran calidad, sin problemas de iluminación, con todos los locutores mirando directamente a la cámara, sin apenas ruido…etc. Esto puede ser un problema a la hora de probar nuestro sistema en condiciones más realistas. Por ello, una línea futura sería la utilización de bases de datos grabadas en condiciones más realistas, con el fin de que el sistema final tuviese mejores prestaciones en el mundo real.
- Otro problema que tiene la base de datos utilizada es el sesgo que posee en edad, sexo y nacionalidad, por ejemplo. Los cuatro locutores de la base de datos son hombres, de mediana edad y británicos, lo que podría hacer que nuestro sistema no funcionase bien con mujeres o niños, o con personas de otros países. Por ello una línea futura es utilizar una base de datos con mayor pluralidad.

- En este trabajo, la extracción de características de la voz se ha basado en el empleo de los coeficientes cepstrales y la frecuencia fundamental, pero existen numerosas características más que pueden dar información sobre el estado emocional del locutor. En una línea futura se deberían investigar éstas otras características, comparando los resultados con los obtenidos en este trabajo.
- Por último, sería de gran interés realizar una aplicación real con el conocimiento adquirido. Esta aplicación podría utilizar la cámara de un teléfono móvil, junto con el micrófono para reconocer el estado emocional del usuario. Este conocimiento podría ser luego utilizado por otras aplicaciones del móvil, o aplicaciones web, para ofrecer experiencias adaptadas al estado emocional del usuario.

Anexo A: English Summary.

Chapter 1: Introduction.

Motivation

In the last decade, the amazing progress of technology has had an important impact on the life of millions of people around the world. Everyday new smart devices are launched to the market and will reach millions of homes in the next years. According to the latest "Mobility Report" by Ericsson, in 2016 the number of mobile lines reached the world population [2]. This leads to more and more humans that interact with machines on a daily basis. In this context, there has been a significant progress in fields such as Natural Language Processing, Automatic Speech Recognition, Speech-To-Text or Speech Synthesis. The main purpose of the research in these areas is to make human-machine interaction more natural. However, these investigations tend to focus on understanding "what they say", namely the content of the message, but not on "how they say it", forgetting the importance of nonverbal communication, for instance, when expressing emotions. Emotions have a crucial role in the human communication process and, therefore, should be taken into account if we aim to build machines able to persuade us, motivate us or "feel" some kind of empathy.

In conclusion, the motivation of this project is to make human-machine interaction more "human" and natural making machines learn how to recognise human emotions.

Objective

The main objective of this project is developing an automatic emotion recognition system based on image and voice.

Furthermore, this project also aims to:

- study the nature of emotions, as well as the different theories that classify them;
- study the basis of Machine Learning, specifically, the task of classification;
- investigate the vocal expression of emotions and the voice features that provide relevant information about it;
- investigate the facial expression of emotions and the visual features that provide relevant information about it;
- evaluate whether the fusion of the voice component and the facial component improves the performance of each component alone.

Applications

Being able to recognize the emotions of people is a truly valuable tool. It is for this reason that this system could be employed in multiple and diverse fields, such as Medicine, Education, Marketing and Retail, Politics or Entertainment, among others.

Previous Works

This project has been based on the research made by the *Centre of Vision, Speech and Signal Processing* (CVSSP) of the University of Surrey in United Kingdom.

Specifically, we have taken as reference the paper "*Speaker-Dependent Audio-Visual Emotion Recognition*" written by Sanaul Haq and Philip J.B. Jackson.

Chapter 2: State of the art.

The nature of the emotions

Many researchers, throughout the last decades, have faced the problem of defining emotions and finding out their origin. As a result, several theories have arisen: evolutive theories, physiological theories, neurological theories and cognitive theories.

Regardless of their origin, they all agree on the existence of three main components in an emotion:

- The subjective experience, which we normally call "feeling"
- The physiological response, which is related to the body reaction.
- The behavioural or expressive response, which is expressed through the facial expressions, the pitch of the voice, its loudness, the gestures…

In order to carry out this project, we have focused on the behavioural or expressive response, since it is the one that allows us to infer the emotional state of a person, based on the observation of his/her facial expression and the analysis of his/her voice.

Emotion classification

The subjective component of emotions makes it difficult to classify them. Regarding this matter, several theories have emerged:

- Basic emotions. According to these theories, there exists a set of basic and universal emotions, such as anger, disgust, happiness or sadness, among others.

- Positive and negative emotions. According to these theories, emotions can be divided in positive emotions, as love, joy or confidence; and negative emotions, as fear, disgust, or sadness.

- Primary and secondary emotions. According to these theories, there are just some primary emotions and the rest of them are secondary emotions that are

derived from the primaries. For instance, happiness would be a primary emotion while euphoria, amusement or enthusiasm would be secondary emotions derived from happiness.

- Representation of emotions in a three-dimensional space. According to these theories, emotions can be represented as a point in a three-dimensional space. For instance, in the PAD model, the axes of this three-dimensional space are: arousal, activation and pleasure.

Machine Learning

Machine Learning is a field in Artificial Intelligence, whose main purpose is providing machines with the capability to learn.

Throughout the learning process of machines we can clearly distinguish two phases: the training and the testing. Below, the reader can find the steps to follow in each phase.

· Training:

1. Acquisition of training instances or observations.
2. Labelling of training instances.
3. Feature extraction.
4. Model or discriminatory function generation.

· Testing:

1. Acquisition of testing instances or observations.
2. Labelling of testing instances.
3. Feature extraction.
4. Classification.
5. Performance evaluation.

Databases

The creation of databases that represent spontaneous emotions can involve several problems. Because of that, most databases used to create emotion recognition systems are acted.

The main advantage of this method is that it allows a greater control over the design of the database. Nevertheless, it has some drawbacks as well:

- Typically, the emotions shown are more exaggerated than the natural ones;
- The system is evaluated under optimal conditions (noiseless audio, high-quality video, good illumination…). However, these conditions are unlikely in real scenarios;

- The categories chosen tend to be basic emotions, but actually they are far more complex.

Chapter 3: Recognition Systems.

General scheme

The final system contains the basic blocks of any classification system:

- A database which contains the data to train and test the recognition system.
- Features extraction techniques to transform each observation from the database into a point in a n-dimensional space.
- A classification algorithm that generates a model which fits the training data and classifies the test data to be evaluated.

Database

The database chosen for this project consists of audio and video recordings of 4 male british-native actors performing 7 different emotions (anger, disgust, happiness, fear, sadness, surprise and neutral). For each emotion, 15 utterances were recorded, except for the neutral emotion for which 30 utterances were recorded. This makes a total of 480 English utterances (120 per actor).

Feature extraction

Feature extraction from images comprises the mean and standard deviation of the coordinates of 60 blue markers painted on the speaker's face during the recording.

Feature extraction from voice is based on Mel frequency cepstral coefficients (MFCCs) and and the fundamental frequency:

- MFCCs is one of the most used types of features in Automatic Speech Recognition (ASR). The main steps to follow when calculating these coefficients are:
    - Fast Fourier Transform (FFT)
    - Mel-frequency spectrum.
    - Logarithm.
    - Inverse Discrete Cosine Transform (IDCT).
- If a voice signal is represented in the time domain, it can be proved that the tonal or voiced sounds (vowels, semi-vowels, and nasals) follow a certain periodic pattern. The frequency of these patterns is known as fundamental frequency.

Classification

The classifiers that have been employed in this project are the Gaussian Mixture Model (GMM) and the Support Vector Machines (SVM) with a Radio Basis Function (RBF) kernel.

Chapter 4: Experiments, results and conclusions.

In the process of implementation, 6 different emotion recognition systems have been generated. They can be classified according to the nature of the data used (image, voice or image and voice) and according to the speaker dependency (speaker-dependent or speaker-independent) being the ultimate system a speaker-independent emotion recognition system based on image and voice.

Different experiments have been conducted in order to analyze the performance of each system with different parameters.

The results of these experiments show that: speaker-dependent systems perform better than speaker-independent ones, as was expected; systems based on image get better results than systems based on voice; the fusion of image features with voice features has better results than each one separated; and the SVM classifier achieves a greater recognition rate than the GMM one.

In conclusion, the main objective of this project has been achieved. The speaker-dependent emotion recognition system based on images and voice (Partial system 5) has got a really impressive performance (98.58%). On the other hand, the speaker-independent system (Ultimate system), despite of its lower recognition rate (76.88%) because of its greater complexity and the insufficient size of the database, has met the expectations.

Chapter 5: Planning and budget.

The project has been developed in a period of 32 weeks, specifically from January 26, 2017 to September 26 2017, with a period of 4 weeks of holidays.

Regarding to the budget, it has been calculated a total project budget of eleven thousand two hundred and ninety-three euros without VAT.

Referencias

[1] *Las ventas de «smartphones» crecen en un 5% en 2016*. Disponible: http://www.abc.es/tecnologia/moviles/telefonia/abci-ventas-smartphones-crecen-5-por-ciento-2016-201702151238_noticia.html [visitado: 08/09/2017].

[2] *El número de líneas móviles alcanza la cifra de habitantes mundiales*. Disponible: http://www.elmundo.es/tecnologia/2016/03/03/56d85088268e3ea0338b4670.html [visitado: 08/09/2017].

[3] Carolina Ferrer Caballero, "8.400 millones de dispositivos conectados a IoT en 2017" Disponible: https://blogthinkbig.com/8-400-millones-de-dispositivos-estaran-conectados-a-internet-a-finales-de-2017 [visitado: 08/09/2017].

[4] *Enganchados al móvil: España, 5º país del mundo que más tiempo pasa con el teléfono. Noticias de Tecnología*. Disponible: https://www.elconfidencial.com/tecnologia/2017-05-26/movil-uso-exceso-espana-salud-enganchados-smartphone_1389117/ [visitado: 08/09/2017].

[5] A. Y. Mor, B. V. Y. Mor and B. Verbal. *The Future of Human-Machine Interaction: It's Not What You Say, It's How You Say It*. Disponible: https://www.wired.com/insights/2014/02/future-human-machine-interaction-say-say/ [visitado: 08/09/2017].

[6] A. Mehrabian and S. R. Ferris, "Inference of attitudes from nonverbal communication in two channels," *J. Consult. Psychol.*, vol. 31, *(3)*, pp. 248-252, 1967.

[7] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch and M. Wróbel, '"Emotion Recognition and Its Applications", vol. 300, 2014.

[8] "Ley orgánica 15/1999, de 13 de diciembre, de protección de datos de carácter personal," vol. I. Disposiciones generales, de 14 de diciembre de, 1999.

[9] "Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos)." 4 de mayo de, 2016.

[10] S. Haq and Philip J.B. Jackson, "Speaker-Dependent Audio Visual Emotion Recognition." , University of Surrey, UK, 2011.

[11] Surrey Audio-Visual Expressed Emotion (SAVEE) Database. Disponible: http://kahlan.eps.surrey.ac.uk/savee/ [visitado: 11/09/2017].

[12] K. Cherry. *What Are the 6 Major Theories of Emotion?*. Disponible: https://www.verywell.com/theories-of-emotion-2795717 [visitado: 11/09/2017].

[13] C. Darwin, *The Expression of the Emotions in Man and Animals.* University of Chicago Press, 1965.

[14] D. Hockenbury and S. E. Hockenbury, *Discovering Psychology.* Worth Publishers, 2016.

[15] K. Cherry. *What Are Emotions and the Types of Emotional Responses?*. Disponible: https://www.verywell.com/what-are-emotions-2795178 [visitado: 11/09/2017].

[16] K. Cherry. *How Many Emotions Are There?*. Disponible: https://www.verywell.com/how-many-emotions-are-there-2795179 [visitado: 11/09/2017].

[17] P. Ekman, W. V. Friesen and P. Ellsworth, *Emotion in the Human Face: Guide-Lines for Research and an Integration of Findings.* Pergamon Press, 1972.

[18] I. Pico, "La rueda de las emociones, de Robert Plutchik," 2016. Disponible: http://psicopico.com/la-rueda-las-emociones-robert-plutchik/ [visitado: 12/09/2017].

[19] A. Mehrabian and J. A. Russell, *An Approach to Environmental Psychology.* Cambridge, MA, US: The MIT Press, 1974.

[20] K. R. Scherer and P. Ekman, *Approaches to Emotion.* Taylor & Francis, 2014.

[21] P. Ekman and W. V. Friesen, *Facial Action Coding System: Manual.* Consulting Psychologists Press, 1978*(1-2).*

[22] *Children's lying behavior towards personified robots: an experimental study*. Disponible: https://www.researchgate.net/figure/297379405_fig8_Figure-9-Facial-action-units-Retrieved-from [visitado: 20/09/2017].

[23] Pedro Isasi Viñuela, Jesús González Boticario and Daniel Borrajo Millán, *Aprendizaje Automático.* (1ª ed.) Sanz y Torres, S.L., 2006.

[24] B. Schuller, G. Rigoll and M. Lang, "Hidden markov model-based speech emotion recognition," in 2003.

[25] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery,* vol. 2, *(2),* pp. 121-167, 1998.

[26] T. L. Nwe, S. W. Foo and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication,* vol. 41, *(4),* pp. 603-623, 2003.

[27] C. M. Bishop, *Neural Networks for Pattern Recognition.* 1995.

[28] W. Wang, *Machine Audition.* (1st ed.) 2011.

[29] *Mel-Frequency Cepstral Coefficients*. Disponible: http://recognize-speech.com/feature-extraction/mfcc [visitado: 18/09/2017].

[30] *Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What's In-Between*. Disponible: http://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html [visitado: 18/09/2017].

[31] C. Busso, Sungbok Lee and S. Narayanan, "Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection," *Tasl*, vol. 17, *(4)*, pp. 582-596, 2009.

[32] "Principal Component Analysis 4 Dummies: Eigenvectors, Eigenvalues and Dimension Reduction," 2013. Disponible: https://georgemdallas.wordpress.com/2013/10/30/principal-component-analysis-4-dummies-eigenvectors-eigenvalues-and-dimension-reduction/. [visitado: 25/09/2017].

[33] A. Ghose. *Support Vector Machines Tutorial*. Disponible: https://blog.statsbot.co/support-vector-machines-tutorial-c1618e635e93 [visitado: 25/09/2017].

[34] *Fundamental Frequency and the Glottal Pulse*. Disponible: https://msu.edu/course/asc/232/study_guides/F0_and_Glottal_Pulse_Period.html [visitado: 20/09/2017].

[35] *How-to simulate Support Vector Machine (SVM) in R*. Disponible: http://en.proft.me/2014/04/22/how-simulate-support-vector-machine-svm-r/ [visitado: 25/09/2017].

[36] *Reconocimiento de firmas off-line mediante máquinas de vectores de soporte*. Disponible: https://www.researchgate.net/figure/277221318_fig1_Fig-2-Transformacion-de-los-datos-de-entrada-a-un-espacio-de-mayor-dimension [visitado: 25/09/2017].